Lecture I: Basic properties of Lorentzian manifolds

1. *Singularity theorems assert the geodesic (in)completeness of certain Lorentzian manifolds.* In Riemannian geometry, the Hopf-Rinow theorem tells us that geodesic completeness is equivalent to the completeness of the manifold as a metric space under the Riemannian distance function:

$$d(p,q) := \inf\{L(\gamma) \ : \ \gamma \text{ piecewise smooth curve segment from } p \text{ to } q\}.$$

Can we replicate this on a Lorentzian manifold, and prove something analogous to Hopf-Rinow? *Unfortunately not, as we'll see below.* Having said that, the attempt to find an analogue of distance on a Lorentzian manifold will lead us to the very important concept of *global hyperbolicity.* So let's start by examining the Hopf-Rinow theorem. (Henceforth all manifolds are connected.)

2. For background on the Riemannian distance function, see [LeeISM, p. 337-341].

3. *Fundamental fact about Riemannian geodesics: they are locally minimizing.* For convenience, let us trace the steps that lead to this result. Let $(\mathcal{U},(x^i))$ be a normal neighborhood centered at p, diffeomorphic to an open set $\mathcal{V}\subset\mathcal{E}_p\subset T_pM$ centered at 0_p, under the exponential map $\exp_p\colon \mathcal{E}_p\longrightarrow M$.

   - *The radial vector field $\partial/\partial r$.* For any $q\in\mathcal{U}-\{p\}$, define the vector field

$$\frac{\partial}{\partial r}\colon\ \mathcal{U}-\{p\}\longrightarrow TM \qquad , \qquad q\ \mapsto\ \left.\frac{\partial}{\partial r}\right|_q = \frac{x^i(q)}{r(q)}\left.\frac{\partial}{\partial x^i}\right|_q, \tag{1}$$

   where $r\colon\mathcal{U}-\{p\}\longrightarrow\mathbb{R}$ is the smooth function $r(q)=\sqrt{(x^1(q))^2+\cdots+(x^n(q))^2}$. It's clear that $\partial/\partial r$ is smooth because its representation in normal coordinates is

$$(x^1(q),\ldots,x^n(q))\ \mapsto\ (x^1(q),\ldots,x^n(q),x^1(q)/r(q),\ldots,x^n(q)/r(q)).$$

   - *$\partial/\partial r$ has unit length in the g-norm and its integral curves are precisely the unit speed radial geodesics $\gamma_V(t)=exp_p(tV)$ for all unit tangent vectors $V\in T_pM$.* To begin with, note that any $q\in\mathcal{U}-\{p\}$ can be written in the form $q=\exp_p(tV)$ for some *unit* vector $V\in T_pM$ and $t\le 1$ (if $V\notin\mathcal{E}_p$, then $t<1$), because

$$q\in\mathcal{U}-\{p\}\Rightarrow q=\gamma_W(1),\ \text{some } W\in\mathcal{E}_p\Rightarrow q=\gamma_{|W|_g\widehat{W}}(1)=\exp_p(|W|_g\widehat{W}),$$

   where $\widehat{W}=W/|W|_g$. And because $\mathcal{V}_p\subset\mathcal{E}_p$ is star-shaped about 0_p, the ray $\{t\widehat{W}:0\le t\le|W|_g\}$ is contained in $\mathcal{V}$. So for any unit $V\in T_pM$ the portion of γ_V that does lie in $\mathcal{U}$ is given by $\gamma_V(t)=\exp_p(tV)$, which in normal coordinates is $\gamma_V(t)=(tV^1,\ldots,tV^n)$ with tangent vector $\gamma_V'(t)=V^i\partial_i|_{\gamma_V(t)}$. Moreover, since $|\gamma_V'(t)|_g=\text{const.}$, we can evaluate it at p, where $g_{ij}(p)=\delta_{ij}$; thus we get $\sum_{i=1}^n(V^i)^2=1$. But we can't evaluate $\partial/\partial r$ at p, and since

$$\left\langle\left.\frac{\partial}{\partial r}\right|_q,\left.\frac{\partial}{\partial r}\right|_q\right\rangle_g=\frac{x^i(q)x^j(q)\,g_{ij}(q)}{(x^1(q))^2+\cdots+(x^n(q))^2},$$

   it's not clear from this that $\langle\partial/\partial r,\partial/\partial r\rangle_g\equiv 1$ (remember that we're not guaranteed that $g_{ij}(q)=\delta_{ij}$ when $q\in\mathcal{U}-\{p\}$). Instead we proceed as follows: setting $q=\gamma_V(t)$ for some unit $V\in T_pM$,

$$\underbrace{q=\gamma_V(t)=\gamma_{tV}(1)\ \mapsto\ \exp_p^{-1}(q)=tV\ \mapsto\ (tV^1,\ldots,tV^n)}_{\text{image of } q \text{ in normal coordinates } (V \text{ need not be in } \mathcal{E}_p)}\ \Rightarrow\ x^i(q)=tV^i$$

$$\begin{aligned}\Rightarrow\quad \left.\frac{\partial}{\partial r}\right|_q &= \frac{tV^i}{\sqrt{\sum_{i=1}^n(tV^i)^2}}\left.\frac{\partial}{\partial x^i}\right|_q\\ &= V^i\left.\frac{\partial}{\partial x^i}\right|_q\\ &= (\gamma_V)'(t).\end{aligned}$$

Hence $\partial/\partial r$ has unit length in the g-norm (and the Euclidean norm, as its definition in eqn. (1) makes evident), and its integral curves are precisely the unit speed radial geodesics $\exp_p(tV) = \gamma_V(t)$.

- *The Gauss lemma: $\partial/\partial r$ is g-orthogonal to the "constant radius" hypersurfaces in $\mathcal{U}$ (so-called "geodesic spheres"), and consequently* $\operatorname{grad} r = \partial/\partial r$ *on* $\mathcal{U} - \{p\}$. These hypersurfaces are defined by diffeomorphically mapping spheres in $\mathcal{V}$ via $\exp_p$ into $\mathcal{U}$, where the radii R of these spheres are established by the norm on T_pM defined by g. Denoting these spheres in $\mathcal{V}$ by $B_R(0_p)$, their image $S_R := \exp_p(B_R(0_p))$ is then a hypersurface in $\mathcal{U}$. Hence if $q \in \mathcal{U}$, so that $q = \exp_p(V)$ for some $V \in \mathcal{V}$, then $|V|_g := R \Rightarrow q \in S_R$. The Gauss lemma thus says that for each $q \in \mathcal{U} - \{p\}$, we have $\langle \partial/\partial r|_q, X_q\rangle_g = 0$ for all $X_q \in T_qS_R \subset T_qM$ (see [LeeRM, p. 102-4] for a proof.) *A nice consequence of the Gauss lemma is that if $q \in S_R$, then any vector $V \in T_qM$ can be orthogonally decomposed as $V = a\,\partial/\partial r + X_q$ for some $X_q \in T_qS_R$.*

(§) *One then uses this decomposition to show that the radial geodesic from p to q is the unique shortest curve from p to q in all of M* (see [LeeRM, p. 105-6]).

- *Corollary 1: every geodesic is locally minimizing.* Let $\gamma\colon I \longrightarrow M$ be a geodesic with $I \subset \mathbb{R}$ open. Whenever a piece $\gamma|_{[t_1,t_2]}$ is contained in a convex open set (and we can always arrange for this to happen, by taking $t_1 < t_0 < t_2$ small enough and picking a convex open set containing $\gamma(t_0)$), then we know that $\gamma|_{[t_1,t_2]}$ *must* be the radial geodesic from $\gamma(t_1)$ to $\gamma(t_2)$, which by (§) is the unique shortest curve connecting these two points.
- *Corollary 2: every unit speed minimizing curve is necessarily a geodesic.* If $\gamma\colon [a,b] \longrightarrow M$ is a minimizing curve (which means that it is minimizing for *any* two of its points), then as before pick any piece $\gamma|_{[t_1,t_2]}$ of it that is contained in a convex open set. Since it is the shortest curve connecting $\gamma(t_1)$ and $\gamma(t_2)$, by (§) it *must* be the radial geodesic segment connecting them. Thus γ solves the geodesic equation in a neighborhood of each of its points, and hence must be a geodesic. (One can also obtain this result using the first variation of arc length; see [LeeRM, p. 100-01].)

4. *Hopf-Rinow theorem:* for a connected Riemannian manifold (M, g), the following are equivalent:

   (1) M is complete as a metric space under Riemannian distance d_g.

   (2) There exists a point $p \in M$ such that $\exp_p$ is defined on all of T_pM.

   (3) M is geodesically complete.

   Remark. The crucial (and most difficult step) in the proof of the Hopf-Rinow theorem is the following: if (2) holds for some $p \in M$, then there is a minimizing geodesic segment from p to any other point q (see [LeeRM, p. 109-11]). Note that (2) is asserted at only one point, although, generally speaking, M is geodesically complete $\Leftrightarrow$ exp is defined on the whole of the tangent bundle TM (see the proof of (3) $\Leftrightarrow$ (2) below).

   Proof of Hopf-Rinow theorem: (3) $\Leftrightarrow$ (2) : for any $p \in M$ and $V \in T_pM$, if the geodesic γ_V exists for all time, then $\gamma_V(1)$ exists, hence $V \in \mathcal{E}_p$. Conversely, for any $p \in M$ and $V \in T_pM$, consider the geodesic γ_V. If $T_pM = \mathcal{E}_p$, it follows that $tV \in \mathcal{E}_p$ for all t, hence $\exp_p(tV) = \gamma_{tV}(1) = \gamma_V(t)$ implies that the latter is defined for all t.

   (1) $\Rightarrow$ (3) : suppose that M is complete as a metric space, and let $\gamma\colon [0,b) \longrightarrow M$ be a unit speed geodesic segment (with b finite), so that it is parametrized by arc length (if $|\gamma'|_g = c \neq 1$, then reparametrize via $\tilde{\gamma}\colon [0, bc) \xrightarrow{\times c^{-1}} [0,b) \longrightarrow M$ so that $|\tilde{\gamma}'|_g = |\gamma'|_g/c = 1$, in which case the arc length function $s(t) = \int_0^t |\tilde{\gamma}'(u)|_g du = t$). We'll now extend γ past b:

   - Given any sequence $\{t_i\} \to b$, set $q_i = \gamma(t_i)$. Since γ is parametrized by arc length,

$$d_g(q_i, q_j) \;\leq\; \underbrace{L(\gamma|_{[t_i,t_j]}) = |t_j - t_i|}_{\text{arc length parametrization is key}} \;\Rightarrow\; \{q_i\} \text{ is Cauchy } \Rightarrow\; \underbrace{\{q_i\} \to q}_{M \text{ is complete}} .$$

- Now that we have a solid "endpoint" q to work with, we're essentially done. Let $\mathcal{C}$ be a convex neighborhood of q contained in a δ-geodesic ball of each of its points. Noting that *every geodesic ball in M is also a metric ball of the same radius* (see [O'Neill, p. 106]),

$$\begin{aligned}\gamma(t_i) = q_i \in \mathcal{C} \quad &\Rightarrow \quad d_g(q_i, q) < \delta \\ &\Rightarrow \quad \text{all geodesics starting at } q_i \text{ exist at least for time } \delta \\ &\Rightarrow \quad \text{the geodesic } \sigma \text{ starting at } q_i \text{ with } \sigma'(0) = \gamma'(t_i) \text{ exists for time } \delta \\ &\Rightarrow \quad \sigma \text{ must be a reparametrization of } \gamma \\ &\Rightarrow \quad \text{if we pick our } t_i \text{ such that } b - \delta < t_i < b, \text{ then } \gamma \textit{ must go past } b.\end{aligned}$$

(3) $\Rightarrow$ (1) : if M is geodesically complete, then exp is defined on the whole of TM. By the remark above, fix any p and let $\{q_i\}$ be a Cauchy sequence in M. Then (2) says that there is a minimizing (unit speed) geodesic segment γ_{V_i} from p to each q_i, which implies that each q_i is in the range of the exponential map: $q_i = \gamma_{V_i}(t) \Rightarrow q_i = \gamma_{tV_i}(1) \Rightarrow tV_i \in \mathcal{E}_p = T_pM$. Hence each of these geodesics can be written in the form $\gamma_{V_i}(t) = \exp_p(tV_i)$. Letting $d_i = d(p, q_i)$, we can also write $q_i = \exp_p(d_i V_i)$ (because the γ_{V_i} are minimizing, hence $L(\gamma_{V_i}) = d(p, q_i)$). Now, since $\{q_i\}$ is a Cauchy sequence, it is bounded, so that $d(q_i, q_j) \leq M$ for all i, j. This implies that $\{d_i\}$ is bounded in $\mathbb{R}$,

$$\underbrace{d(p, q_j)}_{d_j} + \underbrace{d(q_j, q_i)}_{\leq M} \;\geq\; \underbrace{d(p, q_i)}_{d_i} \Rightarrow M \geq d_i - d_j,$$

which in turn implies that the sequence $\{d_i V_i\} \subset T_pM$ is bounded in the g-norm: $|d_i V_i|_g = d_i \leq M$. Being bounded in a compact set (the sphere of radius M), it must have a convergent subsequence $\{d_{i_k} V_{i_k}\} \to V \in T_pM$. By the continuity of the exponential map, we must therefore have $q_{i_k} = \exp_p(d_{i_k} V_{i_k}) \to \exp_p V$. But the original sequence $\{q_i\}$ was Cauchy, so it, too, must converge to the same limit. $\square$

5. *In particular, M compact $\Rightarrow$ M complete $\Rightarrow$ M geodesically complete.* (There is a subtle point here: "M compact" means "compact in the manifold topology," whereas "M complete" means "complete in the metric topology." These two topologies are the same (see [LeeISM, p. 339]), so M will necessarily be compact in the metric topology as well; indeed, if M is compact, then it is complete as a metric space no matter *what* Riemannian metric was used to make it into one.) *But this is not true for Lorentzian manifolds!* A counterexample is provided by the *Clifton-Pohl torus:*

   - Consider $\mathbb{R}^2 - \{0\}$ with the metric

$$\tilde{g} = \frac{1}{u^2 + v^2}\,(du \otimes dv + dv \otimes du). \tag{2}$$

     The only nonzero Christoffel symbols are $\Gamma^u_{uu} = -\frac{2u}{u^2+v^2}$ and $\Gamma^v_{vv} = -\frac{2v}{u^2+v^2}$, and hence the geodesic differential equations are

$$u'' = \frac{2u}{u^2 + v^2}\,(u')^2 \qquad , \qquad v'' = \frac{2v}{u^2 + v^2}\,(v')^2.$$

     It's easy to verify that the curve

$$\gamma\colon (-\infty, 1) \longrightarrow \mathbb{R}^2 - \{0\} \qquad , \qquad t \;\mapsto\; \gamma(t) = \left(\frac{1}{1-t}\,, 0\right) \tag{3}$$

     is a maximally extended geodesic, but obviously *incomplete.*

   - Next, note that the smooth map $\mu\colon (u, v) \mapsto (2u, 2v)$ is an isomety that leaves $\tilde{g}$ unchanged: $\mu^*\tilde{g} = \tilde{g}$. It generates a discrete group of isometries $G = \{\mu^n : n \in \mathbb{Z}\}$, hence a (zero-dimensional)

Lie group. The action of G on $\mathbb{R}^2 - \{0\}$ is *smooth, free* (every isotropy group is trivial), and *proper* (an action $G \times M \longrightarrow M$ is proper if the map $G \times M \longrightarrow M \times M$ defined by $(g, p) \mapsto (g \cdot p, p)$ is proper; if the action of a discrete Lie group is free and "properly discontinuous," then it's proper; see [LeeISM, p. 548]). Since this is the case, the orbit space $T := (\mathbb{R}^2 - \{0\})/G$ is a smooth manifold of dimension $\dim(\mathbb{R}^2 - \{0\}) - \dim G = 2$, and such that the quotient map $\pi\colon (\mathbb{R}^2 - \{0\}) \longrightarrow (\mathbb{R}^2 - \{0\})/G$ is a smooth normal covering map with deck transformation group G (see [LeeISM, p. 549] or [O'Neill, p. 188]). This ensures that there is a unique metric g on T such that $\pi^* g = \tilde{g}$ (see [LeeRM, p. 27]). Since π is a local diffeomorphism, it follows that *the incomplete geodesic given in eqn.* (3) *descends to an incomplete geodesic in* T.

- But this is troubling, because T is *compact:* observe that the action of G reduces $\mathbb{R}^2 - \{0\}$ to an annulus of inner and outer radii 1 and 2, and then further identifies only those points on the inner and outer radii (in particular, each two that lie on a radial line from the origin and intersect these radii), so that T is a torus.
- Note that $\tilde{g}$ (hence g) is *not Riemannian:* for example, the nonzero vector field ∂_u has *zero length,* $\langle \partial_u, \partial_u \rangle_{\tilde{g}} \equiv 0$, and hence so does the tangent vector to $\gamma\colon \langle \gamma', \gamma' \rangle_{\tilde{g}} \equiv 0$. The geodesic γ is what is called a *null* geodesic (in fact *all* null geodesics of $\mathbb{R}^2 - \{0\}$ and the Clifton-Pohl torus T are incomplete; see [O'Neill, p. 260, Exercise 12]).

6. Now that we know the Hopf-Rinow theorem does not hold for non-Riemannian manifolds, we turn to investigating the latter properly, beginning with the definition of a *spacetime.* By a *spacetime* we shall always mean a *connected, time-oriented, 4-dimensional manifold equipped with a Lorentzian metric:*

   - *Connected:* different connected components can't talk to each other, so let's pretend there's only one (physically, we're interested in knowing which events can influence (or be influenced by) a given event, and this requires *curves* joining one point to another).
   - *Lorentz metric:* start with a real finite-dimensional vector space V, and define a *Lorentz product* g to be *a nondegenerate symmetric bilinear form of index* 1, where the *index* of g is defined to be the largest dimension of a subspace $W \subset V$ on which $g|_W$ is *negative definite.* Here are some properties of Lorentz products:

     (a) Vectors $v \in V$ now fall into three classes:

     $$\begin{cases} \text{spacelike} & \Leftrightarrow \quad \langle v, v \rangle_g > 0 \text{ or } v = 0, \\ \text{timelike} & \Leftrightarrow \quad \langle v, v \rangle_g < 0, \\ \text{null} & \Leftrightarrow \quad \langle v, v \rangle_g = 0 \text{ and } v \neq 0. \end{cases}$$

     Timelike and null vectors are collectively called *causal vectors.* Note that 0 is not causal.

     (b) *Orthogonal complements:* given a subspace $W \subset V$, define as usual the subspace

     $$W^\perp = \{v \in V \ : \ v \perp W\},$$

     but notice that you cannot call this the "orthogonal complement of W" because in general $W + W^\perp$ is *not all of* V. Here's an easy example: in *Minkowski spacetime* $\mathbb{R}^2_1 = (\mathbb{R}^2, g)$, with g defined by $\langle v, w \rangle_g = v^1 w^1 - v^2 w^2$, take $W = \operatorname{span}((1, 1))$; then $W^\perp = W$! (Notice that W is a degenerate subspace, and this is no coincidence.) Having said that, it is always true that

     $$\dim W + \dim W^\perp = \dim V \qquad , \qquad (W^\perp)^\perp = W, \tag{4}$$

     and that $g|_W$ *is nondegenerate* $\Leftrightarrow V = W \oplus W^\perp$ (see [O'Neill, p. 49]). (Note that $(W^\perp)^\perp = W$ implies that W is nondegenerate $\Leftrightarrow W^\perp$ is nondegenerate.) If $g|_W$ is a Lorentz product, we say that W is a *timelike subspace;* if $g|_W$ is degenerate, we say that W is a *lightlike subspace* (e.g., $\operatorname{span}((1, 1)) \subset \mathbb{R}^2_1$ is a lightlike subspace); if $g|_W$ is positive definite (i.e., an inner product space), we say that W is a *spacelike subspace.* Using eqn. (4) one can show that

     $$W \text{ is timelike} \Leftrightarrow W^\perp \text{ is spacelike},$$

     so that W is lightlike $\Leftrightarrow W^\perp$ is lightlike (see [O'Neill, p. 141]).

(c) *Orthonormal bases:* an easy induction argument shows that *every Lorentz vector space has an orthonormal basis,* where we now understand a vector v to be orthonormal $\Leftrightarrow \langle v, v\rangle_g = \pm 1$ (see [O'Neill, p. 50]). Given an orthonormal basis $\{e^1, \ldots, e^n\}$ for V, we define its signature to be $(\varepsilon^1, \ldots, \varepsilon^n) := (\langle e^1, e^1\rangle_g, \ldots, \langle e^n, e^n\rangle_g)$. With an orthonormal basis in hand, one can show that *the signature of any orthonormal basis for V is an invariant, equal to the index of g on V, namely 1* (see [O'Neill, p. 51]).

(d) *Orthogonal causal vectors are necessarily null and collinear* (see [O'Neill, p. 155, Exercises 2(b)]). Using this one can prove the following two important facts:

- W timelike with $\dim W \geq 2 \Rightarrow W$ contains two linearly independent null vectors,
- W lightlike $\Rightarrow W$ contains no timelike vectors and its nullcone is contained in a one-dimensional subspace.

(See [O'Neill, p. 141-2] for proofs.)

(e) *Definition of Lorentzian metric:* a *Lorentzian metric* g on a smooth manifold M is a *symmetric nondegenerate* $(0,2)$ *tensor field of constant index 1,* hence a Lorentz product $g_p \colon T_pM \times T_pM \longrightarrow \mathbb{R}$ at each $p \in M$.

(f) *Not all manifolds admit Lorentzian metrics!* All noncompact smooth manifolds do, but compact manifolds admit Lorentzian metrics only if their Euler characteristic is zero; so, for example, the only compact 2-manifolds that admit Lorentzian metrics are the torus, Möbius band, and the Klein bottle (but not $\mathbb{S}^2$, genus g surfaces with $g \geq 2$, or the real projective plane). Generally speaking, *a smooth manifold admits a Lorentzian metric $\Leftrightarrow$ it admits a nonvanishing global vector field* (one direction is easy: ($\Leftarrow$) suppose that M has a nonvanishing vector field $W \in \mathfrak{X}(M)$; pick any Riemannian metric g on M (all smooth manifolds admit Riemannian metrics, using a simple partition of unity argument; see [LeeISM, p. 329]) and consider now the unit vector field $\widehat{W} := W/|W|_g$. Letting $\omega := g(\widehat{W}, \cdot)$ denote the 1-form metrically equivalent to $\widehat{W}$, define a new metric on M by

$$g_L := g - 2\,\omega \otimes \omega.$$

This metric will be Lorentzian (for example, at any $p \in M, \langle \widehat{W}_p, \widehat{W}_p\rangle_{g_L} = 1 - 2 = -1$). For the ($\Rightarrow$) direction, see [O'Neill, p. 149] and the reference therein.)

- *Time orientation:* roughly speaking, a "continuous" choice of "past" and "future" direction at each tangent space. To understand what is meant by this, first recall the concept of *orientation* for any smooth manifold: at each $p \in M$, we partition all ordered bases in T_pM according to whether the change of basis matrix from one (ordered) basis to another has positive determinant. Under this equivalence relation T_pM has precisely two equivalence classes, and we *orient* M by simply picking one equivalence class at each T_pM. To have a notion of "continuity of choice," we say that M is *smoothly oriented* if at each point we can find a smooth local frame that has the given orientation at each point of its domain (see [LeeISM, p. 380]).

  Though it is completely independent of orientation, the concept of *time orientation* works the same way. Instead of ordered bases at each T_pM we have *timecones:* namely, for any timelike vector $u \in T_pM$ we define the *timecone of T_pM containing u* to be

$$C(u) = \{v \text{ any timelike vector in } T_pM \ : \ \langle u, v\rangle_{g_p} < 0\}.$$

  The fundamental fact regarding timecones is that two timelike vectors $u, v \in T_pM$ are in the same timecone $\Leftrightarrow \langle u, v\rangle_{g_p} < 0$ (see [O'Neill, p. 143]; the proof relies heavily on the fact that $T_pM = \text{span}(u) \oplus u^\perp$ for any timelike $u \in T_pM$, which follows from (b) above). Hence

$$u \in C(v) \Leftrightarrow v \in C(u) \Leftrightarrow C(u) = C(v) \Rightarrow \{\text{timelike vectors in } T_pM\} = C(u) \,\dot\cup\, C(-u).$$

  (Notice also that timecones are *convex:* $v, w \in C(u) \Rightarrow av + bw \in C(u)$ whenever $a \geq 0, b \geq 0$ are not both zero.) So just as with orientation, we can *time-orient* our manifold by picking one of two

timecones from each tangent space T_pM. Analogous to local frames above, we then say that a given time orientation is *smooth* provided that for any $p \in M$ we can find a neighborhood $\mathcal{U}$ of p and a timelike vector field $X_{\mathcal{U}} \in \mathfrak{X}(\mathcal{U})$ such that each $X_{\mathcal{U}}|_p \in T_pM$ is in the chosen timecone. Now imagine if we had a smooth *global* timelike vector field $\tau \in \mathfrak{X}(M)$; then assigning to each $p \in M$ the timecone containing τ_p would smoothly time-orient M. What's particularly nice is that the other direction is also true: if one had a smooth time orientation in place already, then one can piece together all the local timelike vector fields $X_{\mathcal{U}}$ into one smooth global timelike vector field, using a partition of unity $\{\psi_\alpha\}$ subordinate to the cover $\{\mathcal{U}_\alpha\}$; the smooth composite vector field $\sum_\alpha \psi_\alpha X_\alpha$ will then be timelike and have the given time orientation at every point (you'll need the fact that timecones are convex; see [O'Neill, p. 145]). Thus we have that *a Lorentz manifold is time-orientable* $\Leftrightarrow$ *there exists a smooth timelike vector field* $\tau \in \mathfrak{X}(M)$. (Notice that this is analogous to the existence of a nonvanishing n-form for a manifold to be orientable.) Given a time orientation τ, the timecones in which each τ_p lives will be called *future timecones.* The other timecone in each tangent space will be called the *past timecone.*

- *"Backwards" Cauchy-Schwarz and triangle inequalities:* For any two timelike vectors v, w in a Lorentz vector space, $|\langle v, w\rangle| \geq |v||w|$; and if v and w are in the same timecone, then $|v| + |w| \leq |v + w|$, with equality $\Leftrightarrow$ v, w are collinear (see [O'Neill, p. 144]).

7. *Piecewise smooth causal curves:* the classifications "spacelike," "timelike," and "null" extend in an obvious way to spacelike, timelike, and null piecewise smooth curves: for example, a piecewise smooth curve is *spacelike* if its tangent vector at every point along it is spacelike. Having said that, the timelike and null cases require care. In particular, we say that a piecewise smooth curve α in a spacetime M is *timelike* provided that each $\alpha'(t)$ is timelike, *and* that at each "kink" t_i

$$\langle \alpha'(t_i^-), \alpha'(t_i^+)\rangle_g < 0,$$

where the first vector derives from $\alpha|_{[t_{i-1},t_i]}$ and the second from $\alpha|_{[t_i,t_{i+1}]}$, both of which are smooth. In other words, α' is not permitted to switch timecones at a break (in particular, timelike curves cannot be going "forward in time" and then suddenly stop and go "backward in time"). A piecewise smooth timelike curve α is *future-pointing* provided its tangent vector always lies in the future timecone (given a time orientation $\tau \in \mathfrak{X}(M)$, this means that $\langle \alpha'(t), \tau_{\alpha(t)}\rangle_{g_{\alpha(t)}} < 0$ at every point along α.

What about piecewise smooth null curves? Since their tangent vectors are null, not timelike, they don't live in timecones. To be able to say whether they are future- or past-pointing, we thus expand our notion of timecone to include null vectors by defining what are called *causal cones:* the causal cone of a timelike vector u is defined to be the set $\overline{C}(u)$ of all causal vectors w such that $\langle v, w\rangle_g < 0$. (One can show that causal vectors v, w are in the same causal cone $\Leftrightarrow$ either $\langle v, w\rangle_g < 0$ or v and w are both null with $w = av, a > 0$; one can also show that $\overline{C}(u) =$ closure of $C(u) - 0$; see [O'Neill, p. 155, Exercise 3]). We can then speak of future and past causal cones. The obvious example of a causal cone is the causal cone of any point in Minkowski spacetime, which is the timecone at that point plus its boundary. Given a time orientation $\tau \in \mathfrak{X}(M)$, we say that a piecewise smooth null curve γ is *future-pointing* provided that $\langle \gamma'(t), \tau_{\gamma(t)}\rangle_{g_{\gamma(t)}} < 0$, or equivalently, that $\gamma'(t) \in \overline{C}(\tau_{\gamma(t)})$ at every point $\gamma(t)$ along γ. Finally, we can speak of future-pointing causal curves, which are piecewise smooth curves each of whose smooth segments is either future-pointing timelike or future-pointing null.

8. *In Minkowski spacetime* $\mathbb{R}^4_1$, *"future-pointing" ("past-pointing") really does mean "forward in time" ("backward in time"), and timelike curves really do correspond to speeds less than that of light.* Recall that for any $p = \big(x^0(p), x^1(p), x^2(p), x^3(p)\big) \in \mathbb{R}^4_1$, we have the orthonormal basis

$$\{\partial_0|_p, \partial_1|_p, \partial_2|_p, \partial_3|_p\} \subset T_p\mathbb{R}^4_1 \qquad , \qquad \langle \partial_0, \partial_0\rangle_g = -1 \qquad , \qquad \langle \partial_i, \partial_i\rangle_g = 1 \qquad (i = 1, 2, 3).$$

The timelike vector field ∂_0 thus time-orients $\mathbb{R}^4_1$, so if α is any a future-pointing timelike curve in $\mathbb{R}^4_1$, then by definition $\alpha'(s)$ and $\partial_0|_{\alpha(s)}$ are in the same timecone:

$$\langle \alpha'(s), \partial_0|_{\alpha(s)}\rangle < 0 \Rightarrow \frac{d(x^0 \circ \alpha)}{ds}(s) = \frac{d\alpha^0}{ds}(s) > 0 \Rightarrow \alpha^0(s) \text{ is monotonically increasing.} \tag{5}$$

Hence α is always "going forward" in the time coordinate x^0: in fact α can never pass through the same t-constant hypersurface twice. Moreover, if α has domain a connected interval $I \subset \mathbb{R}$, then eqn. (5) says that α^0 is a diffeomorphism mapping I onto some interval $J \subset \mathbb{R}$; let $u\colon J \longrightarrow I$ denote its inverse. This defines a curve

$$\vec{\alpha}\colon J \longrightarrow \mathbb{R}^3 \qquad , \qquad t \in J \;\mapsto\; \vec{\alpha}(t) = \underbrace{\left(\alpha^1(u(t))\,,\, \alpha^2(u(t))\,,\, \alpha^3(u(t))\right)}_{t = \alpha^0(s) \Rightarrow \alpha^i(u(t)) = \alpha^i(s)},$$

which by the chain rule satisfies

$$\frac{d\vec{\alpha}}{dt} = \frac{d\vec{\alpha}}{ds}\frac{ds}{dt}.$$

The function $\vec{\alpha}$ is called the *associated Newtonian particle* of the curve α, with $|d\vec{\alpha}/dt|$ its *speed;* this is the actual speed that an observer would measure (see [O'Neill, p. 167-8]). To show that this speed must be less than unity, simply observe that because $\langle \alpha'(s), \alpha'(s) \rangle < 0$,

$$\alpha'(s) \;=\; \underbrace{\frac{d\alpha^0}{ds}}_{\alpha^0(s)=t}\partial_0|_{\alpha(s)} + \sum_{i=1}^{3}\frac{d\alpha^i}{ds}\,\partial_i|_{\alpha(s)} \Rightarrow \left|\frac{dt}{ds}\right| > \left|\frac{d\vec{\alpha}}{ds}\right| \Rightarrow \left|\frac{d\vec{\alpha}}{dt}\right| < 1.$$

If α were a *null* geodesic, then eqn. (5) would still hold. In that case $\langle \alpha'(s), \alpha'(s) \rangle \equiv 0$, so we would get $|d\vec{\alpha}/dt| = 1$; *in other words, light travels at the speed of light!*

(§§) *In stark contrast to the Riemannian case* (§) *above, timelike radial geodesics in a Lorentz manifold maximize length in normal neighborhoods:* in any normal neighborhood $\mathcal{U}$ of a point p in a Lorentzian manifold, *if* there exists a *timelike* curve from p to a point $q \in \mathcal{U} - \{p\}$, then the radial geodesic segment from p to q is the *unique longest* timelike curve in $\mathcal{U}$ from p to q (see [O'Neill, p. 147]; the length of a timelike curve segment $\alpha\colon [a,b] \longrightarrow M$ is given similarly to the Riemannian case, $L(\alpha) := \int_a^b \sqrt{|\langle \alpha'(u), \alpha'(u) \rangle_g|}du$, the only difference being the absolute value sign. Note that the term "length" in Lorentzian geometry can be misleading, since null curves have zero length). *But this forces us to ask: is it possible to have a suitable notion of distance on an arbitrary spacetime? The answer is yes, but to see how, we must first go back to Minkowski spacetime.*

9. *Separation of points in Minkowski spacetime.* Note that $\mathbb{R}^4_1$ is itself a (global) normal neighborhood of each of its points, hence radial geodesic segments exist for all time; *by* (§§) *the (timelike) line segments are the unique longest curves between any two points that can be connected by a timelike curve.*

   Consider now any "freely falling material particle" in $\mathbb{R}^4_1$. Such particles travel on timelike geodesics, so consider a future-pointing timelike geodesic $\alpha\colon [0,l] \longrightarrow \mathbb{R}^4_1$ from p to q, parametrized by arc length, or *proper time* τ, so that α has unit speed and the length (the "elapsed proper time") is $L(\alpha) = l$. Since we know that α is a line segment from p to q,

$$\alpha(\tau) = \left((1 - (\tau/l))\, x^0(p) + (\tau/l)\, x^0(q)\,,\, (1 - (\tau/l))\, \vec{x}^{\,i}(p) + (\tau/l)\, \vec{x}^{\,i}(q)\right),$$

   so that

$$\begin{aligned} 1 \;\equiv\; |\alpha'(\tau)| \;&=\; \sqrt{\left|\langle \alpha'(\tau), \alpha'(\tau) \rangle\right|} \\ &=\; \left| -(\alpha^0)'(\tau)^2 + \sum_{i=1}^{3}(\alpha^i)'(\tau)^2 \right|^{1/2} \\ &=\; \underbrace{\frac{1}{l}\left| -\left((x^0(q) - x^0(p)\right)^2 + \sum_{i=1}^{3}\left((x^i(q) - x^i(p)\right)^2 \right|^{1/2}}_{|\vec{pq}| = l}. \end{aligned} \tag{6}$$

The number $|\vec{pq}| = l$ in eqn. (6) is known as the *separation* between p and q; it is of course equal to $L(\alpha)$, the length or elapsed proper time of the freely falling material particle from p to q (in the context of special relativity, a freely falling spaceship records $|\vec{pq}|$ as the time from event p to event q). In terms of the exponential map, $\exp_p(\vec{pq}) = q$, where $\vec{pq} = \sum_{i=0}^{3}(x^i(q) - x^i(p))\,\partial_i$. *Now we modify our question above: can we use the notion of separation to determine a distance function on an arbitrary spacetime? It turns out that we can, but before proceeding to definitions let's first define chronological futures and pasts in Lorentzian manifolds.*

10. *Chronological futures and pasts.* Let M be a spacetime. For any $p, q \in M$, define the following *causality relations on M:*

    (a) $p \ll q \Leftrightarrow$ there is a future-pointing *timelike* curve in M from p to q,

    (b) $p < q \Leftrightarrow$ there is a future-pointing *causal* curve in M from p to q.

    Note that all causal curves are by definition piecewise smooth, and are not allowed to jump timecones or causal cones during a "kink" (see (7) and (8) above). Since timelike curves are an example of causal curves, evidently $p \ll q \Rightarrow p < q$. We also write $p \leq q$ to indicate that either $p < q$ or $p = q$ (note that if there are no causal curves from p to itself (no "closed" causal curves), then $p \not< p$, because the constant curve $\alpha(p) \equiv p$ has zero tangent vector, hence is not causal.) Now we generalize these causality relations: for any subset $A \subset M$, define

    (a) *Chronological future:* $I^+(A) = \{q \in M \ : \text{ there is a } p \in A \text{ with } p \ll q\}$,

    (b) *Chronological past:* $I^-(A) = \{q \in M \ : \text{ there is a } p \in A \text{ with } q \ll p\}$,

    (c) *Causal future:* $J^+(A) = \{q \in M \ : \text{ there is a } p \in A \text{ with } p \leq q\}$,

    (d) *Causal past:* $J^-(A) = \{q \in M \ : \text{ there is a } p \in A \text{ with } q \leq p\}$.

    Any open set $\mathcal{U} \subset M$ is a spacetime in its own right; if we wish to restrict our causality relations to $A \subset \mathcal{U}$, we write, for example, $I^+(A,\mathcal{U})$ to denote the chronological future of A in the manifold $\mathcal{U}$; likewise for the other causal sets. Certainly $I^+(A,\mathcal{U}) \subset I^+(A) \cap \mathcal{U}$, but note that in general $I^+(A,\mathcal{U}) \neq I^+(A) \cap \mathcal{U}$. Here are some properties of chronological and causal future sets:

    - $A \cup I^+(A) \subset J^+(A)$,
    - $I^+(A) = \bigcup_{p\in A} I^+(p)$ and $J^+(A) = \bigcup_{p\in A} J^+(p)$ (note that in general $p \notin I^+(p)$, but $p \in J^+(p)$)

    (*i*) $x \ll z \Rightarrow$ there are infinitely many points y such that $x \ll y \ll z$ (similarly for $\leq$),

    (*ii*) $\begin{cases} x \ll y \text{ and } y \leq z \Rightarrow x \ll z, \\ x \leq y \text{ and } y \ll z \Rightarrow x \ll z, \end{cases}$ (see [O'Neill, p. 294]; the proof is a variational result),

    - $I^+(A) \overset{(i)}{=} I^+(I^+(A)) \overset{(ii)}{=} I^+(J^+(A)) \overset{(ii)}{=} J^+(I^+(A)) \subset J^+(J^+(A)) \overset{(i)}{=} J^+(A)$,
    - $I^\pm(A)$ *is an open set for any subset* A (see [O'Neill, p. 403] for a proof, which relies heavily on the fact that for any convex open set $\mathcal{C}$ containing p, the set $I^+(p,\mathcal{C})$ is open).

    Example 1: Minkowski spacetime. The set $I^+(p)$ is just the future timecone of p, and $J^+(p) = \overline{I^+(p)}$ is the causal cone. This suggests an easy relationship between chronological and causal sets, but it is misleading. Indeed, *causal sets J^+ need not be closed:* consider Minkowski spacetime $\mathbb{R}^2_1$ with the point $(1,1)$ deleted (see [O'Neill, p. 404, Figure 1]). Then $\overline{I^+(\mathbf{0})} \neq J^+(\mathbf{0})$ in $\mathbb{R}^2_1 - \{(1,1)\}$ and the latter is not closed, since any point (n,n) with $n > 1$ is a limit point of $J^+(\mathbf{0})$ but is not contained in $J^+(\mathbf{0})$. Having said that, for any spacetime M and any subset $A \subset M$,

    $$\operatorname{int} J^+(A) = I^+(A) \quad , \quad J^+(A) \subset \overline{I^+(A)}, \text{ with equality} \Leftrightarrow J^+(A) \text{ is closed}$$

    (see [O'Neill, p. 404] for a proof).

Example 2: the Lorentz cylinder. Consider the product manifold $(\mathbb{S}^1_1 \times \mathbb{R}, g = -d\varphi \otimes d\varphi + dt \otimes dt)$. Causality here is as trivial as possible: $I^+(p) = \mathbb{S}^1_1 \times \mathbb{R} = J^+(p)$ for all p (see [O'Neill, p. 148, 402]).

Crucial fact: variational results establish that if α is a future-pointing causal curve from a set A to a point $q \in J^+(A) - I^+(A)$, *then α is necessarily a null geodesic that has no conjugate points before q and does not meet $I^+(A)$* (picture any point on a nullcone in Minkowski spacetime). For a proof, see [O'Neill, p. 298, 404]. *Hence for any subset A, the set $J^+(A)$ is a union of $A, I^+(A)$, and (possibly) certain null geodesics from A. This fact play will play a crucial role in Penrose's singularity theorem.*

"Time travel" is possible on compact spacetimes: cover your spacetime M by the open sets $\{I^+(p)\}_{p\in M}$. By compactness there is a finite subcover among these, say $I^+(p_1), \ldots, I^+(p_k)$. If $I^+(p_1)$ is contained in any other $I^+(p_i)$'s, discard it. This forces $p_1 \in I^+(p_1)$, because in general $p \in I^+(A) \Rightarrow I^+(p) \subset I^+(A)$. Hence M contains a closed future-pointing timelike curve. Since we don't want this to happen on physical grounds, we dismiss compact spacetimes from consideration, and designate a spacetime to satisfy the *chronology condition* if it contains no closed timelike curves, and the *causality condition* if it contains no closed causal curves. Clearly the causality condition $\Rightarrow$ the chronology condition, but the converse is not true. Here is an example of a Lorentzian manifold that satisfies the latter but not the former: consider the smooth action of the discrete Lie group $\mathbb{Z}$ on $\mathbb{R}^2_1$ given by $n\cdot(t,x) = (t+n, x+n)$. As with the Clifton-Pohl torus above, this action is free and "properly discontinuous," hence proper; thus the orbit space $\mathbb{R}^2_1/\mathbb{Z}$ is a smooth (two dimensional) manifold with the quotient map $\pi\colon \mathbb{R}^2_1 \longrightarrow \mathbb{R}^2_1/\mathbb{Z}$ a smooth covering map. Null geodesics $s \mapsto (s,s)$ in $\mathbb{R}^2_1$ project to closed null geodesics $\mathbb{R}^2_1/\mathbb{Z}$, but no timelike curves do (any closed timelike curve in $\mathbb{R}^2_1/\mathbb{Z}$ would lift to a timelike curve from (t,x) to $(t+n, x+n)$, which is impossible).

11. *Lorentzian distance function:* recall that our goal is to generalize the notion of separation of points in Minkowski spacetime, as defined in eqn. (6) above. To that end, for any spacetime M and points $p, q \in M$, define their *time separation*

$$\boldsymbol{\tau}(p,q) := \sup\{L(\gamma) \ : \ \gamma \text{ future-pointing causal curve segment from } p \text{ to } q\}.$$

Here are some basic properties of the time separation $\boldsymbol{\tau}$:

(a) If the set of lengths between p and q is unbounded, then set $\boldsymbol{\tau}(p,q) = \infty$.

(b) Recalling our crucial fact above, it's clear that $\boldsymbol{\tau}(p,q) > 0 \Leftrightarrow q \in I^+(p)$. If $q \notin J^+(p)$, then $\boldsymbol{\tau}(p,q) = 0$, but the converse is in general *not* true: recalling our crucial fact above, if $q \in J^+(p) - I^+(p)$, then $\boldsymbol{\tau}(p,q) = 0$.

(c) *In Minkowski spacetime, $\boldsymbol{\tau}(p,q) = |\vec{pq}|$ if $p \leq q$ and is otherwise zero, exactly as we want.*

(d) *Time separation behaves badly if the chronology condition fails:* $p \in I^+(p) \Rightarrow \boldsymbol{\tau}(p,p) = \infty$ (just keep riding the same future-pointing timelike curve). In $\mathbb{S}^1_1 \times \mathbb{R}$, $\boldsymbol{\tau}(p,p) = \infty$ for all p.

(e) *In fact time separation can behave very badly:* for a fixed $p \in M$, if $\boldsymbol{\tau}(p,q) = \infty$ for all $q \in M$, then $\boldsymbol{\tau}(p,q) = \infty$ for *all points* $p, q \in M$ (see [BEE, p. 137]).

(f) *Time separation is in general not symmetric:* $\boldsymbol{\tau}(p,q) \neq \boldsymbol{\tau}(q,p)$. (Note that if you ride a future-pointing causal curve "backwards," then you will be *past-pointing:* for example, if $\alpha\colon [0,b] \longrightarrow M$ is a future-pointing timelike curve from $\alpha(0) = p$ to $\alpha(b) = q$, then $\tilde{\alpha}(t) := \alpha(b-t)$ is a past-pointing timelike curve from $\tilde{\alpha}(0) = q$ to $\tilde{\alpha}(b) = p$, since $\langle \tilde{\alpha}'(t), \alpha'(t)\rangle = \langle -\alpha'(t), \alpha'(t)\rangle > 0$.)

(g) *"Reverse" triangle inequality:* $p \leq q \leq r \Rightarrow \boldsymbol{\tau}(p,q) + \boldsymbol{\tau}(q,r) \leq \boldsymbol{\tau}(p,r)$. (Proof: if there is no future-pointing causal curve from say p to q, then $\boldsymbol{\tau}(p,q) = 0$, so that $p \leq q \Rightarrow p = q$, so the result clearly holds. So assume that there are future-pointing causal curves α from p to q and β from q to r. If that's the case, then for any $\delta > 0$ there must exist curves α and β such that $\boldsymbol{\tau}(p,q) - \delta/2 < L(\alpha)$ and $\boldsymbol{\tau}(q,r) - \delta/2 < L(\beta)$, respectively, so that

$$\boldsymbol{\tau}(p,r) \geq L(\alpha+\beta) = L(\alpha) + L(\beta) > \boldsymbol{\tau}(p,q) + \boldsymbol{\tau}(q,r) - \delta.$$

Since $\delta > 0$ was arbitrary, we're done.)

(h) *The time-separation function* $\boldsymbol{\tau}: M \times M \longrightarrow [0, \infty)$ *is not in general continuous, but it is always lower semicontinuous* (see [O'Neill, p. 410-11]).

(i) *Properties that ensure the existence of a longest unbroken causal geodesic between two points* p *and* q *with* $p < q$. Begin by constructing a family $\{\alpha_n\}$ of future-pointing causal curve segments from p to q whose lengths $L(\alpha_n)$ converge to $\boldsymbol{\tau}(p, q)$ (this is easy: for each n, there must exist a piecewise smooth causal curve α_n satisfying $L(\alpha_n) > \boldsymbol{\tau}(p, q) - 1/n$). Notice that each of these curves is in $J^+(p) \cap J^-(q) := J(p, q)$. *Suppose now this set is compact and that your spacetime is causal: then these two conditions suffice to ensure that there will be a causal (generally broken) geodesic* λ *from* p *to* q *with* $L(\lambda) = \boldsymbol{\tau}(p, q)$ (this is a key result in Lorentzian geometry and depends on some subtle analysis; see [O'Neill, p. 409] and [Bernal-Sanchéz07]). But we can do better: we can also ensure that λ must be *unbroken,* as follows. If λ is an unbroken null geodesic, then $L(\lambda) = \boldsymbol{\tau}(p, q) = 0 \Rightarrow q \in J^+(p) - I^+(p)$. Otherwise, there is a (generally piecewise smooth) timelike curve from p to q arbitrarily close to λ (see [O'Neill, p. 294]), which implies that λ cannot have any null pieces, since in such a case the timelike curve will have strictly longer length, a contradiction, since $L(\lambda) = \boldsymbol{\tau}(p, q)$. So if λ is not an unbroken null geodesic, then it must be a (possibly broken) timelike geodesic. But if it is broken, then for any fixed-endpoint variation of λ, the fact that $L(\lambda) = \boldsymbol{\tau}(p, q)$ means that λ will be a critical point of the first variation of arc length, hence λ must be an unbroken geodesic; see [O'Neill, p. 265].

(j) *In conclusion, we may say that if a spacetime satisfies the causality condition and the sets* $J(p, q) = J^+(p) \cap J^-(q)$ *are compact for all* $p < q$*, then there is an unbroken causal geodesic from* p *to* q *with length* $\boldsymbol{\tau}(p, q)$*. Such a spacetime is called globally hyperbolic.* This is very nice: in a globally hyperbolic spacetime, any pair of points that can be joined by a causal curve can be joined by a longest unbroken causal geodesic. Minkowski spacetime, Robertson-Walker spacetimes, and Kruskal spacetime are all globally hyperbolic. (Notice, however, that removing a single point from a globally hyperbolic spacetime destroys the property, for then there will be noncompact $J(p, q)$'s.) As we'll soon see, there's an easier way to determine whether a spacetime is globally hyperbolic.

(k) *On a globally hyperbolic spacetime* M*, the Lorentzian distance function* $\boldsymbol{\tau}$ *is continuous and finite-valued* (see [O'Neill, p. 412] and [BEE, p. 140]).

Lecture II: Cauchy hypersurfaces

1. *The easy argument in eqn.* (5) *above works in any spacetime in which "the flow of time is geodesic" and "initially normal to something," in a sense that we make precise below.* (The motivation behind this is given by the following question: what really makes the argument leading to eqn. (5) work? Is it because the time orientation of Minkowski spacetime $\mathbb{R}^4_1$ happens to *coincide* with a (global) coordinate vector field, $\partial/\partial x^0$? Not really: by the flow box theorem (see the proof below), we can *always* make the time orientation coincide (locally) with a coordinate vector field of some coordinate chart. If not that, then is it the *orthogonality* of the ∂_i's in $\mathbb{R}^4_1$? While this certainly made the conclusion in eqn. (5) follow easily, it turns out that we can weaken even *this* assumption and still derive the same conclusion. As we now show, what *really* enables the ease of argument in eqn. (5) is that the *integral curves of the time orientation are geodesics; to put it more bluntly: the flow of time is geodesic in* $\mathbb{R}^4_1$.)

   Indeed, consider a spacetime (M, g) with time orientation $T \in \mathfrak{X}(M)$. At a point $p \in M$, assume the following two conditions:

   (a) The integral curves of T are all geodesics in a neighborhood of p, in which case we'll also assume that T is a unit vector field (if not, then simply consider T/c in place of T, where $c = \sqrt{-\langle T, T\rangle_g}$, so that $\langle T/c, T/c\rangle_g \equiv -1$),

   (b) T is orthogonal to a certain spacelike hypersurface Σ_0 containing p that we define below.

   By the flow box theorem, there exists a connected coordinate chart $(\mathcal{U}, (x^i))$ centered at p with respect to which $T = \partial/\partial x^0$ (this has nothing to do with the metric!). Now, by (a) the integral curves of T are geodesics in a neighborhood of p, so we may assume that they are geodesics everywhere in $\mathcal{U}$ (possibly after shrinking $\mathcal{U}$). Let $\theta^{(p)} \colon \mathcal{D}^{(p)} \longrightarrow \mathcal{U}$ denote the geodesic integral curve of T starting at p (here $\mathcal{D}^{(p)} \subset \mathbb{R}$ is a connected open interval containing 0). Next, consider the $x^0 = 0$ slice

$$\Sigma_0 := \{q \in \mathcal{U} \ : \ x(q) = \big(0, x^1(q), x^2(q), x^3(q)\big)\}. \tag{7}$$

   This is obviously a hypersurface in M that separates $\mathcal{U}$ into two connected components; of course, $p \in \Sigma_0$ and $T_q\Sigma_0 = \operatorname{span}(\partial_1|_q, \partial_2|_q, \partial_3|_q)$ for all $q \in \Sigma_0$. *Let's assume that T is normal to $\Sigma_0 \cap \mathcal{U}$, so that $\langle \partial_0|_q, \partial_i|_q\rangle_g = 0$ for all $q \in \Sigma_0$.* Then because $(\theta^{(q)})'(t) = \partial_0|_{\theta^{(q)}(t)}$, the derivative of $\langle \partial_0, \partial_i\rangle_g$ along any $\theta^{(q)}$ must be zero:

$$\frac{d}{dt}\langle \partial_0, \partial_i\rangle_g = \langle \underbrace{\nabla_{\partial_0}\partial_0}_{0}, \partial_i\rangle_g + \langle \partial_0, \nabla_{\partial_0}\partial_i\rangle_g = -\langle \partial_0, \nabla_{\partial_i}\partial_0\rangle_g = -\frac{1}{2}\nabla_{\partial_i}\underbrace{\langle \partial_0, \partial_0\rangle_g}_{-1} = 0 \Rightarrow \langle \partial_0, \partial_i\rangle_g = \text{const.}$$

   In particular, if $q \in \Sigma_0$ then $\langle \partial_0|_q, \partial_i|_q\rangle_g = 0$, so this constant is zero. *In other words, $\partial_0 \perp \partial_i$ on all of $\mathcal{U}$.* In fact it's easy to show that $T|_{\mathcal{U}} = -\operatorname{grad} x^0$, the latter of course being automatically normal to Σ_0 (see [O'Neill, p. 106]). It's enough to show that $\operatorname{grad}(T, \cdot) = -dx^0$; indeed, for all $X \in \mathfrak{X}(\mathcal{U})$,

$$\operatorname{grad}(T, \cdot)(X) = \langle \partial_0, X\rangle_g = X^0\langle \partial_0, \partial_0\rangle_g = -X^0 = -dx^0(X).$$

   (This immediately implies that T is orthogonal to each x^0-constant slice Σ_t, with Σ_t being defined analogously to eqn. (7) with t in place of 0). Now the argument in eqn. (5) carries over identically: if α is any future-pointing timelike curve passing through $\mathcal{U}$, then

$$\langle \alpha'(s), \partial_0|_{\alpha(s)}\rangle < 0 \ \Rightarrow \ \frac{d\alpha^0}{ds}(s) > 0 \ \Rightarrow \ \alpha^0(s) \text{ is monotonically increasing,}$$

   so that α can never pass a "time slice" Σ_t more than once. Hence the two conditions (a) and (b) above are sufficient to ensure that "future-pointing" really does mean "forward in time." *In other words, just as with Minkowski spacetime above, α can never pass through the same t-constant hypersurface twice.* "Time slices" of the form Σ_t seem to be quite special submanifolds indeed. We now devote some time to generalizing them.

2. *Cauchy surfaces* are modeled after t-constant hyperplanes in $\mathbb{R}^4_1$. These have the following properties:

   (a) They are hypersurfaces (closed, embedded codimension 1 submanifolds)

   (b) They are *achronal:* the relation $p \ll q$ never holds on them, so no timelike curve meets it more than once; in fact *every* inextendible future-pointing timelike curve meets it *exactly once* (this is exactly what we showed in eqn. (5) above).

   (c) They have no *"edges"* (a notion we make precise below).

   What does (b) say physically? Well, consider a point $p \in \mathbb{R}^4_1$. Every *inextendible* timelike curve through p will meet our t-constant hyperplane (in fact so will any inextendible null curve), so any observer or light signal that can influence a system at p must have gone through this hyperplane at some point. Bottom line: if we have "initial data" on our hyperplane, then the state of any system at p is completely determined by this hyperplane. Indeed, initial data on our hyperplane determines the entire evolution of $\mathbb{R}^4_1$, past and future. (Hence the name "Cauchy surface," for this is analogous to the *Cauchy problem* for PDEs, where one seeks a solution f to a PDE that simultaneously satisfies "initial data" on a hypersurface S, the initial data being given in the form of a smooth function $\varphi\colon S \longrightarrow \mathbb{R}$, so that our solution f will satisfy $f|_S = \varphi$.)

3. Of these, (b) is the most important: in fact it essentially *implies* (a) and (c). Indeed, suppose a subset A of a spacetime M satisfies (b). Then heuristically, we expect:

   - (b) $\Rightarrow$ (c): if A had an "edge," then one can imagine that just past this edge there would be timelike curves that don't meet A at all. (It's easy to make this notion precise: a point $p \in \bar{A}$ is an *edge point* of A if for every neighborhood $\mathcal{U}$ of p there exists a timelike curve from $I^-(p,\mathcal{U})$ to $I^+(p,\mathcal{U})$ that does *not* meet A. Finally, bear in mind that although the concept of edge seems identical to that of boundary, they're not the same; for one, it depends crucially on dimension: think of a finite segment of the x-axis in $\mathbb{R}^2_1$ (two edge points), and the same line viewed as lying in $\mathbb{R}^3_1$ (every point on the line segment is an edge point).)
   - (b) $\Rightarrow$ (a): suppose A was a submanifold; then it couldn't be one- or two-dimensional, as that wouldn't be "big" enough to satisfy (b) (think of curves or surfaces in $\mathbb{R}^4_1$, or to help visually, points and curves in $\mathbb{R}^3_1$); it also couldn't be four-dimensional, as that would be too "big" (think of an open set and a timelike curve going through it).

4. So let's concentrate on (b). First, note that a subset $A \subset M$ is *achronal* $\Leftrightarrow I^+(A) \cap I^-(A) = \varnothing$

   (Proof: ($\Rightarrow$) if $q \in I^+(A) \cap I^-(A)$, then we would have $p_1 \ll q \ll p_2$, hence $p_1 \ll p_2$. Contradiction. ($\Leftarrow$) if we had $p_1 \ll p_2$ for $p_1, p_2 \in A$, then any point on the timelike curve between these points would be in $I^+(A) \cap I^-(A)$. Contradiction.) Just as for t-constant hyperplanes in $\mathbb{R}^4_1$: everything "above" is the future, everything "below" is the past, and the hyperplane neatly divides the two (again, no edges). Second, note that the closure $\bar{A}$ of an achronal set A is necessarily achronal, and $\bar{A} - A \subset \operatorname{edge} A$ (Proof: given any limit point $p \in A$, suppose that $p \ll q$ for some $q \in A$. Let $\mathcal{C}$ be a convex set of p and pick a point $q' \in \mathcal{C}$ that is on the timelike curve joining p and q, so that $p \ll q' \in \mathcal{C}$. Since $p \in I^-(q',\mathcal{C})$ and $I^-(q',\mathcal{C})$ is open, there exists a point $p' \in A \cap I^-(q',\mathcal{C})$. But then $p' \ll q' \ll q$, violating achronality (the case $q \ll p$ is similar, with $I^+(q',\mathcal{C})$ in place of $I^-(q',\mathcal{C})$). To show that $\bar{A} - A \subset \operatorname{edge} A$, let p be a limit point of A not in A; we just need to find, for a given neighborhood $\mathcal{U}$ of p, some timelike curve from $I^-(p,\mathcal{U})$ to $I^+(p,\mathcal{U})$ that doesn't meet A, and this is easy, because *no* timelike curve through p can *ever* meet A. Indeed, if it did, say at some point $q \in A$, then we would have $p \ll q$ or $q \ll p$, and we just showed this can never happen if A is achronal.)

   ($\star$) *[Note: the method "pick a point q' on the curve just before/after $p \in \mathcal{C}$, so that $p \in I^{\pm}(q',\mathcal{C})\ldots$" is a common technique; in fact, this is how one proves that $I^{\pm}(A)$ is open for any set $A \subset M$, after having shown that the sets $I^{\pm}(q',\mathcal{C})$ are open whenever $\mathcal{C}$ is convex; see [O'Neill, p. 403].]*

($\star\star$) *An achronal set $A \subset M$ has no edge points (edge $A = \varnothing$) $\Leftrightarrow$ A is a closed topological hypersurface.* (So an edgeless achronal set *must* have a manifold structure, though not necessarily a *smooth* one, since, for example, any nullcone $\Lambda^+(p) \subset \mathbb{R}^4_1$ is edgeless and achronal.) To prove this, we first need to prove the "flow box" theorem.

5. *Flow box coordinates:* if $T \in \mathfrak{X}(M)$ is nonzero at a point $p \in M$, then there exist coordinates (u^i) about p such that $T = \partial/\partial u^1$.

   - Start by picking smooth coordinates $(\mathcal{U}, \varphi, (x^i))$ centered at p such that in these coordinates $T_p \in T_pM$ has a nonzero x^1-component. Henceforth we consider only $T|_{\mathcal{U}} \in \mathfrak{X}(\mathcal{U})$, though we'll continue to write T. Let $\theta\colon\ \mathcal{D} \longrightarrow \mathcal{U}$ be the flow of T; we will denote the unique maximal integral curve starting at a point $q \in \mathcal{U}$ by $\theta^{(q)}\colon \mathcal{D}^{(q)} \longrightarrow \mathcal{U}$, where $\mathcal{D}^{(q)}$ is a connected open interval in $\mathbb{R}$ containing 0. Since $\mathcal{D} \subset \mathbb{R} \times \mathcal{U}$ is open, pick a basis element $(-\varepsilon, \varepsilon) \times \mathcal{U}_0 \subset \mathcal{D}$ containing $(0, p)$, which maps under $1_{\mathbb{R}} \times \varphi$ to a basis element $(-\varepsilon, \varepsilon) \times \varphi(\mathcal{U}_0)$ of $(0, \mathbf{0})$. Consider also the $x^1 = 0$ slice of $\varphi(\mathcal{U}_0)$, viewed as an open set $S \subset \mathbb{R}^{n-1}$:

   $$S := \{\mathbf{u} = (u^2, \ldots, u^n) : (0, \mathbf{u}) \in \varphi(\mathcal{U}_0)\}.$$

   Now we define a smooth map that will serve as our new coordinates:

   $$\psi\colon\ (-\varepsilon, \varepsilon) \times S \longrightarrow \mathcal{U} \qquad , \qquad \psi(t, \mathbf{u}) = \theta^{(\varphi^{-1}(0,\mathbf{u}))}(t). \tag{8}$$

   What does ψ do? Well, for fixed $\mathbf{u} \in S$, the map $t \mapsto (t, \mathbf{u})$ is an integral curve of $\partial/\partial u^1$ (just a line parallel to the u^1-axis), and this curve is mapped by ψ to the integral curve of T starting at $\varphi^{-1}(0, \mathbf{u})$ (more precisely, to the $(-\varepsilon, \varepsilon)$-portion of it). *In other words, ψ maps integral curves of $\partial/\partial u^1$ to integral curves of T; once we prove that ψ really is a coordinate chart, this will show that in the coordinates (u^i) the integral curves of T are lines parallel to the u^1-axis.* (For starters, note that ψ is centered at p, $\psi(0, \mathbf{0}) = \theta^{(p)}(0) = p$, and that ψ is smooth, because it is the composition

   $$(t, \mathbf{u}) \ \mapsto\ (t, (0, \mathbf{u})) \in (-\varepsilon, \varepsilon) \times \varphi(\mathcal{U}_0) \ \mapsto\ (t, \varphi^{-1}(0, \mathbf{u})) \ \mapsto\ \theta^{(\varphi^{-1}(0,\mathbf{u}))}(t),$$

   all of which pieces are smooth maps.)

   - Another way to think about these coordinates: consider again the $x^1 = 0$ slice of $\varphi(\mathcal{U}_0)$ which, thought of as an open set in $\mathbb{R}^{n-1}$, we called S above. This slice is obviously a hypersurface in $\mathcal{U}_0 \subset M$, and notice that T is nonzero on it: specifically, $T^1(p) \neq 0$, where $p = \varphi^{-1}(0, \mathbf{0})$, so the component T^1 is nonzero in a neighborhood of p, hence nonzero on our slice in $\mathcal{U}_0$, after possibly shrinking $\mathcal{U}_0$ a bit (this also implies that T is not tangent to this hypersurface, because $T^1 \neq 0$ along it). *So our coordinates start out at points $\varphi^{-1}(0, \mathbf{u})$ on this slice, and then ride along the flow of T through that point, for some time t.*
   - Now we just need to show that ψ is a local diffeomorphism about $(0, \mathbf{0})$. To that end, note that the action of $d\psi_{(0,\mathbf{0})}$ on the vectors $\{\partial/\partial u^2|_{(0,\mathbf{0})}, \ldots, \partial/\partial u^n|_{(0,\mathbf{0})}\} \subset T_{(0,\mathbf{0})}((-\varepsilon, \varepsilon) \times S)$ is

   $$d\psi_{(0,\mathbf{0})}\left(\left.\frac{\partial}{\partial u^i}\right|_{(0,\mathbf{0})}\right) = \frac{\partial \hat{\psi}^j}{\partial u^i}(0, \mathbf{0})\left.\frac{\partial}{\partial x^j}\right|_{\psi(0,\mathbf{0})} = \left.\frac{\partial}{\partial x^i}\right|_p \qquad , \qquad i = 2, \ldots, n,$$

   where we remind the reader that since $\psi(0, \mathbf{u}) = \varphi^{-1}(0, \mathbf{u})$, the coordinate representation $\hat{\psi}$ on the slice $\{(0, \mathbf{u})\}$ is simply $\hat{\psi}(0, \mathbf{u}) = (0, \mathbf{u})$, so that $\frac{\partial \hat{\psi}^j}{\partial u^i}(0, \mathbf{u}) = \delta_{ij}$ for $i = 2, \ldots, n$.
   - Next, what about the action of $d\psi_{(t_\circ,\mathbf{u}_\circ)}$ on $\partial/\partial t|_{(t_\circ,\mathbf{u}_\circ)}$, for *any* $(t_\circ, \mathbf{u}_\circ) \in (-\varepsilon, \varepsilon) \times S$? Well, if we fix $\mathbf{u}$, then $\psi(\cdot, \mathbf{u})\colon (-\varepsilon, \varepsilon) \longrightarrow \mathcal{U}$ is just an integral curve of T, so its derivative is obvious. Indeed, for any $f \in C^\infty((-\varepsilon, \varepsilon) \times S)$,

   $$d\psi_{(t_\circ,\mathbf{u}_\circ)}\left(\left.\frac{\partial}{\partial t}\right|_{(t_\circ,\mathbf{u}_\circ)}\right) f = \left.\frac{\partial}{\partial t}\right|_{(t_\circ,\mathbf{u}_\circ)} (f \circ \psi)(t, \mathbf{u}) = \left.\frac{d}{dt}\right|_{t_\circ} \big(f \circ \theta^{(\varphi^{-1}(0,\mathbf{u}_\circ))}\big)(t) = \theta^{(\varphi^{-1}(0,\mathbf{u}_\circ))'}(t_\circ) f,$$

   and of course, the latter is just $T_{\psi(t_\circ,\mathbf{u}_\circ)} f$.

- Putting these two results together, we see that under the action of $d\psi_{(0,\mathbf{0})}$,

$$\{\partial/\partial t|_{(0,\mathbf{0})}, \partial/\partial u^2|_{(0,\mathbf{0})}, \ldots, \partial/\partial u^n|_{(0,\mathbf{0})}\} \mapsto \{T_p, \partial/\partial x^2|_p, \ldots, \partial/\partial x^n|_p\}.$$

  The latter is a basis because T_p has non-zero x^1-component. In other words, $d\psi_{(0,\mathbf{0})}$ takes a basis to a basis, hence is an isomorphism, which means that ψ is a local diffeomorphism about $(0,\mathbf{0})$ and thus a coordinate chart centered at p. But more than that, it's a coordinate chart satisfying $d\psi(\partial/\partial t) = T$ at *all* points in its domain, as we wanted. Renaming t to u^1, we're done.

6. *Proof of* $(\star\star)$. $(\Leftarrow)$ Suppose A is an achronal closed topological hypersurface. First, let's ask: if A has edge points, can it *contain* any of them? *No:* for any $p \in A$, let $\mathcal{U}$ be a slice coordinate chart for A containing p, and assume $\mathcal{U}$ is connected so that $\mathcal{U} - A$ splits into two components. We'd like to argue that $I^-(p,\mathcal{U})$ and $I^+(p,\mathcal{U})$ cannot be in the same component, so that any curve from one to the other *must* go through A. Here's why: first, note that $I^\pm(p,\mathcal{U})$ are open and connected (both by a $(\star)$ argument); furthermore, they are disjoint and don't meet A because A is achronal. To see why they can't live in the same component, pick a connected convex neighborhood $\mathcal{C} \subset \mathcal{U}$ centered at p such that A also splits $\mathcal{C}$ into two components (we could have just let $\mathcal{C} = \mathcal{U}$ from the beginning). Then any timelike radial geodesic through p must meet both components (such a geodesic *must* exist, too; for example, a timelike curve goes through p via the time orientation of M; see [O'Neill, p. 147]), and this immediately implies that at least one of $I^\pm(p,\mathcal{U})$ must lie on both sides of A; but because both are *connected,* they must lie on separate sides, each by itself. So $p \notin \operatorname{edge} A$. Now we ask: can A *have* any edge points to begin with? *No:* by our argument above, we must have $\bar{A} - A \subset \operatorname{edge} A$, but A is closed so $\bar{A} = A$. (Of course, we could have just written this and been done with this part of the proof, but it was important to show that whether or not an achronal set A is closed, if it is a topological hypersurface then it can't contain any edge points to begin with, whether or not it has any.)

   $(\Rightarrow)$ Given $p \in A$, consider the timelike vector field $T \in \mathfrak{X}(M)$ determined by the time orientation of M. It is nonzero at p, so there exist "flow box" coordinates $(\mathcal{U}, (t, \mathbf{u}))$ centered at p in which $T = \partial/\partial t$. Shrinking $\mathcal{U}$ if necessary (for example, by working inside a convex neighborhood $p \in \mathcal{C} \subset \mathcal{U}$), *we may write* $(-\varepsilon, \varepsilon) \times S$ *in eqn.* (8) *in the form* $(-a-\delta, a+\delta) \times S$ *and such that the slices* $\psi(-a \times S)$ *and* $\psi(a \times S)$ *lie in* $I^-(p,\mathcal{U})$ *and* $I^+(p,\mathcal{U})$*, respectively.* (To do this, consider the open connected set $I^+(p,\mathcal{U}) \subset \mathcal{U}$. Pick a point $q \in I^+(p,\mathcal{U})$ along the integral curve $\theta^{(p)}(t)$ through p, and consider $I^+(q,\mathcal{U}) \subset I^+(p,\mathcal{U})$. Within this open connected set, all of whose points are timelike-connected to q and hence to p, pick a small open rectangle whose "upper top" is less than 2ε across (notice that all of these upper top points are timelike-connected to p). Say $\theta^{(p)}(t)$ intersects this upper top at $t = a$. Repeat this procedure for a point $q' \in I^-(p,\mathcal{U})$, and suppose the intersection here takes place at $t = -b$. Now just pick the smaller of a and $|b|$, say a, and shrink the domain of ψ to $(-a-\delta, a+\delta) \times S$ for small δ.)

   For fixed $\mathbf{u} \in S$, consider the integral curve $\theta^{(\varphi^{-1}(0,\mathbf{u}))}(t)$ through $\varphi^{-1}(0,\mathbf{u}) := q \in \mathcal{U}$. Here's where the edgelessness and achronality of A come into play: *this integral curve must meet A or else q is an edge point, because we made sure that $\theta^{(q)}(-a) \in I^-(p,\mathcal{U})$ and $\theta^{(q)}(a) \in I^+(p,\mathcal{U})$. But A is edgeless by assumption, so $\theta^{(q)}(t)$ meets A, and does so only once because A is achronal.* Now we ask ourselves: what is the t coordinate of this meeting point? Well, for each $\mathbf{u} \in S$, the corresponding integral curve $\theta^{(\varphi^{-1}(0,\mathbf{u}))}(t)$ meets A at precisely one value of its parameter t, say $t_\mathbf{u}$. This, of course, defines a function

$$h\colon S \longrightarrow (-a, a) \qquad , \qquad h(\mathbf{u}) = t_\mathbf{u},$$

   which, it turns out, is *continuous* (see [O'Neill, p. 414]), and which in turn defines the following homeomorphism, analogous to the map that sends the nullcone $\Lambda^+(p) \subset \mathbb{R}^4_1$ to a horizontal plane:

$$\phi\colon \mathcal{U} \longrightarrow \mathbb{R}^n \qquad , \qquad \theta^{(\varphi^{-1}(0,\mathbf{u}))}(t) \;\mapsto\; (t, \mathbf{u}) \;\mapsto\; (t - h(\mathbf{u}), \mathbf{u}).$$

   The inverse of the second part is $\phi^{-1}(t,\mathbf{u}) = (t + h(\mathbf{u}), \mathbf{u})$, so ϕ is indeed a homeomorphism. More important is what it does to points in $A \cap \mathcal{U}$:

$$\theta^{(\varphi^{-1}(0,\mathbf{u}))}(h(\mathbf{u})) \;\mapsto\; (0, \mathbf{u}),$$

thus confirming that in the coordinates provided by $(\phi, \mathcal{U})$, the subset $A \cap \mathcal{U}$ is a topological $t = 0$ slice, and hence that A is a topological hypersurface. Finally, A is necessarily closed because $\bar{A} - A \subset \operatorname{edge} A = \varnothing$, as we showed above. $\square$

7. *Closed, achronal topological hypersurfaces are plentiful.* In fact, for any set $A \subset M$, the boundaries of $I^+(A)$ or $J^+(A)$, if either exists, are closed achronal topological hypersurfaces, because both $I^+(A)$ and $J^+(A)$ are *future sets,* in that they contain their chronological futures: $I^+(I^+(A)) \subset I^+(A)$ and $I^+(J^+(A)) \subset J^+(A)$ (Proof: set $F = I^+(A)$ and suppose $\operatorname{bd} F$ exists. Then $I^+(\operatorname{bd} F) \subset F$ and $I^-(\operatorname{bd} F) \subset M - F$, so that $I^+(\operatorname{bd} F) \cap I^-(\operatorname{bd} F) = \varnothing$ and $\operatorname{bd} F$ is achronal (because if $I^+(F) \subset F$, then $I^-(M - F) \subset M - F$). Finally, since $I^+(\operatorname{bd} F) \subset \operatorname{int} F$ and $I^-(\operatorname{bd} F) \subset \operatorname{int}(M - F)$, it's simply impossible to have a timelike curve go from $I^-(p)$ to $I^+(p)$ *without* meeting $\operatorname{bd} F$, for $p \in \operatorname{bd} F$. Hence $\operatorname{bd} F$ cannot have any edge points.)

8. *Cauchy hypersurfaces:* remember, we want these to be modeled after t-constant hyperplanes in $\mathbb{R}^4_1$; i.e., to represent an "instant of time." Recognizing that property (b) above was the crucial one, let us make the following definition: a *Cauchy hypersurface S in a spacetime M is any subset that is met exactly once by every inextendible timelike curve.* This definition has the following consequences:

   - *S is a closed achronal topological hypersurface:* S is clearly achronal, so if we could show that S is also edgeless, then we'd immediately know that it must be a closed topological hypersurface. Actually, we'll just show that S is the boundary of a future set, as follows: M is partitioned into the disjoint sets $I^+(S), I^-(S), S$, and S must be the common boundary of $I^\pm(S)$, because any timelike curve through a point $p \in S$ must immediately meet both $I^-(S)$ and $I^+(S)$, so we're done.
   - In fact S is also met by every inextendible *causal* curve, but the proof is technical; see [O'Neill, p. 415-6]. Some authors, in fact, *define* a Cauchy hypersurface to be an achronal subset that is met exactly once by every inextendible causal curve (see, e.g., [BEE, p. 65]).
   - *The only interesting topology in M takes place in S, because M retracts onto S.* The idea is that one can use the flow of the time orientation of M to map every point uniquely to some point in S, with the points already in S remaining stationary. (Proof: let T be the time orientation of M and consider its flow $\theta\colon \mathcal{D} \longrightarrow M$, where $\mathcal{D} \subset \mathbb{R} \times M$ is an open set containing the slice $0 \times M$. Since $S \subset M$ is a topological hypersurface, $\mathbb{R} \times S \subset \mathbb{R} \times M$ is also a topological hypersurface, hence $\mathcal{D}_S := \mathcal{D} \cap (\mathbb{R} \times S)$ is an open (topological) submanifold of $\mathbb{R} \times S$ that is four-dimensional. But then the continuous map

     $$\theta|_{\mathcal{D}_S}\colon \mathcal{D}_S \longrightarrow M \qquad , \qquad (t, p) \;\mapsto\; \theta^{(p)}(t)$$

     is a bijection between manifolds of the same dimension, hence a homeomorphism. Composing with the projection $\pi_S\colon \mathbb{R} \times S \longrightarrow S$, we then immediately have the retraction we want:

     $$r := \pi_S|_{\mathcal{D}_S} \circ (\theta|_{\mathcal{D}_S})^{-1}\colon M \longrightarrow \mathcal{D}_S \longrightarrow S. \tag{9}$$

     In particular, since our spacetime M is connected, so is S. If we happen to have two Cauchy hypersurfaces S and S', then using their respective retractions as maps from one to the other will prove that they must be homeomorphic.)
   - *Geroch's splitting theorem [Geroch]: if a spacetime M has a Cauchy hypersurface, then M is homeomorphic to $\mathbb{R} \times S$.* (Proof: by above, M is homeomorphic to an open subset $\mathcal{D}_S \subset \mathbb{R} \times S$, but actually, we may as well have taken T to be *complete,* so that $\mathcal{D} = \mathbb{R} \times M$ and consequently $\mathcal{D}_S = \mathbb{R} \times S$. Why? Because if the flow of T is not complete, then consider $T/|T|_{g_R}$ instead, where g_R is any *complete* Riemannian metric on M (that is to say, every Cauchy sequence in (M, g_R) converges; such a metric always exists on any connected smooth manifold; see [LeeISM, p. 346, Problem 13-18]).) Clearly $T/|T|_{g_R}$ determines the same time orientation as T.

- *Not all Cauchy hypersurfaces are smooth!* (For example, in $\mathbb{R}^2_1$ let $\mathcal{C}$ be a t-constant line approaching from the left, then stopping at $x^1 = 0$ and sloping down (linearly) for a short while, then going horizontally again to the right. If the "slant" line has an acute angle less than $45°$ on each side (so that it doesn't lie on a null cone), then $\mathcal{C}$ will be a Cauchy hypersurface, and clearly not a smooth one.)

9. *Globally hyperbolic* $\Leftrightarrow$ *Cauchy hypersurface:* for a proof, which uses volume functions, see [Geroch]. (For more on volume functions, see [Dieckmann].)

10. *Smooth Geroch splittings:* [Bernal-Sanchéz03] recently showed that any globally hyperbolic spacetime M necessarily contains a *smooth* Cauchy hypersurface S such that M is *diffeomorphic* to $\mathbb{R} \times S$.

11. *Exotic* $\mathbb{R}^4$*'s cannot be globally hyperbolic:* [Chernov-Nemirovski] has shown that *a contractible, globally hyperbolic Lorentzian* n*-manifold must be diffeomorphic to* $\mathbb{R}^n$*. In particular, since all the "exotic"* $\mathbb{R}^4$*'s are homeomorphic (but not diffeomorphic) to the standard* $\mathbb{R}^4$*, they are contractible and thus cannot be globally hyperbolic.* The result relies on the smooth splitting of [Bernal-Sanchéz03] and the three-dimensional Poincaré conjecture.

Lecture III: Penrose's singularity theorem

1. *We are only concerned with those manifolds that are as "large as can be."* In particular, define a semi-Riemannian manifold (M, g) to be *extendible* if (M, g) can be isometrically embedded as a proper open submanifold of a connected semi-Riemannian manifold $(\widetilde{M}, \tilde{g})$; in other words, if there is an open submanifold $\widetilde{P} \subset \widetilde{M}$ furnished with the induced metric $\tilde{g}|_{\widetilde{P}}$ and an isometry $\iota\colon (M, g) \longrightarrow (\widetilde{P}, g|_{\widetilde{P}})$. Here are some examples of inextendible manifolds:

   - M *compact* $\Rightarrow$ (M, g) *inextendible,* simply because compact subsets of Hausdorff spaces are closed (in particular, the Clifton-Pohl (2) torus is inextendible).
   - (M, g) *geodesically complete* $\Rightarrow$ (M, g) *inextendible* (Proof: suppose M isometrically embeds as a proper open submanifold of a connected $\widetilde{M}$ and pick a point $p \in \partial M$ (since $\widetilde{M}$ is connected, such a p exists). Pick a convex open set $\widetilde{\mathcal{C}} \subset \widetilde{M}$ of p with radius δ; being open, $\widetilde{\mathcal{C}}$ set must contain a point $q \in M$, so consider the radial unit speed geodesic segment $\tilde{\gamma}$ from p to q in $\widetilde{\mathcal{C}}$. Assume that $\tilde{\gamma}(0) = p$ and $\tilde{\gamma}(t_0) = q$, where $|t_0| < \delta$ ($\tilde{\gamma}$ is unit speed and contained in $\mathcal{C}$, hence its length $L(\tilde{\gamma}) = t_0 < \delta$). Then clearly $\gamma(t) := \tilde{\gamma}(t_0 - t)$ is a geodesic in $M \subset \widetilde{M}$ starting at q that cannot be extended past $t = \delta$, contradicting the geodesic completeness of M).
   - *Even in the Riemannian setting, the class of inextendible manifolds is strictly larger than the class of geodesically complete manifolds.* An example is provided by the two dimensional "positive" null cone $\Lambda^+(\mathbf{0})$ (minus the vertex $\mathbf{0}$) in Euclidean space $(\mathbb{R}^3, \bar{g})$. As a smooth Riemannian manifold, $(\Lambda^+(\mathbf{0}), \bar{g}|_{\Lambda^+(\mathbf{0})})$ is inextendible. Now let $\{q_i\} \subset \Lambda^+(\mathbf{0})$ be any sequence that converges to $\mathbf{0}$ in $\mathbb{R}^3$. Then $\{q_i\}$ is necessarily a Cauchy sequence, but it doesn't converge in $\Lambda^+(\mathbf{0})$. By the Hopf-Rinow theorem, $(\Lambda^+(\mathbf{0}), \bar{g}|_{\Lambda^+(\mathbf{0})})$ is therefore geodesically incomplete.
   - *Kruskal spacetime is inextendible:* in Kruskal spacetime K, the curvature invariant $I := R^{ijkl}R_{ijkl} = 48m^2/r^6$. Now, if Kruskal spacetime were geodesically complete, then it would be inextendible by our discussion above. Of course, we know that it is *not* so: for example, any timelike curve that enters the black hole region ($r < 2m$) necessarily *ends* at a finite parameter interval at the singularity $r = 0$ (see eqn. (15) below, and also [O'Neill, p. 392]). Indeed, one can show that a unit speed timelike geodesic $\gamma\colon [0, b) \longrightarrow K$ in Kruskal spacetime is incomplete $\Leftrightarrow$ $r \circ \gamma(\tau) \to 0$ as $\tau \to b$ (see [O'Neill, p. 396-7]). This means that if we evaluate I along *any* inextendible timelike geodesic γ, it will be unbounded as $\tau \to b$, because $\tau \to b \Rightarrow r \to 0$. *And this in turn implies that Kruskal spacetime must be inextendible.* Indeed, assume that K isometrically embeds as a proper open submanifold of a connected manifold $\widetilde{K}$. Then using the same convex open set argument as above, we can construct an inextendible timelike geodesic $\gamma\colon [0, b) \longrightarrow K$ which *is* extendible when viewed as a geodesic $\tilde{\gamma}$ in $\widetilde{K}$. But clearly $I(\gamma(\tau))$ does *not* have a finite limit as $\tau \to 0$, since $\tau \to 0 \Rightarrow r \circ \gamma(\tau) \to 0$, whereas $I(\tilde{\gamma}(\tau))$ clearly *does* (being well defined at $\tau = b$), and this is a contradiction.

2. *Trapped surfaces in Kruskal spacetime.* In Kruskal spacetime t-constant spheres $\mathbb{S}^2(r)$ of a fixed radius r are points on the Kruskal plane. Consider such a point (really, a closed 2-submanifold of Kruskal spacetime). The metric in Kruskal coordinates is

$$\tilde{g} = F(r)\, du \otimes dv + F(r)\, dv \otimes du + r^2 d\vartheta \otimes d\vartheta + r^2 \sin^2\vartheta\, d\varphi \otimes d\varphi,$$

   where

$$F(r) := \frac{8m^2}{r}\, e^{1-r/2m} \qquad , \qquad \phi(r) := e^{r/2m-1}\,(r - 2m) = uv.$$

   Being a point on the Kruskal plane, the induced metric on $\mathbb{S}^2(r)$ is clearly

$$\tilde{g}|_{\mathbb{S}^2(r)} = r^2 d\vartheta^2 + r^2 \sin^2\vartheta\, d\varphi^2. \tag{10}$$

Now at each $p \in \mathbb{S}^2(r)$, the normal space $(T_p\mathbb{S}^2(r))^\perp$ always contains two linearly independent null vectors (see [O'Neill, p. 141]). Clearly $\operatorname{grad} u = \tilde{g}^{uv}\partial_v$ and $-\operatorname{grad} v = -\tilde{g}^{uv}\partial_u$ are two linearly independent, future-pointing (smooth) null vector fields orthogonal to $\mathbb{S}^2(r)$ (recall that Kruskal spacetime is time oriented by $\partial_v - \partial_u$). The covector fields metrically equivalent to them are

$$\langle \operatorname{grad} u, \cdot \rangle_{\tilde{g}} = du \qquad , \qquad \langle -\operatorname{grad} v, \cdot \rangle_{\tilde{g}} = -dv.$$

The divergence of $\operatorname{grad} u$ is

$$\begin{aligned} \operatorname{div}_{\mathbb{S}^2(r)}(\operatorname{grad} u) &= (\tilde{g}|_{\mathbb{S}^2(r)})^{ij}(\operatorname{grad} u)_{i;j} \\ &= (\tilde{g}|_{\mathbb{S}^2(r)})^{ij}\left[\partial_j(\operatorname{grad} u)_i - (\operatorname{grad} u)_p \widetilde{\Gamma}^p_{ji}\right] \\ &= \tilde{g}^{\vartheta\vartheta}\Big[0 - (\operatorname{grad} u)_u \underbrace{\widetilde{\Gamma}^u_{\vartheta\vartheta}}_{-F^{-1}r\, r_v}\Big] + \tilde{g}^{\varphi\varphi}\Big[0 - (\operatorname{grad} u)_u \underbrace{\widetilde{\Gamma}^u_{\varphi\varphi}}_{-F^{-1}r\, r_v\, \sin^2\vartheta}\Big] \\ &= \frac{2\, r_v}{rF(r)} = \frac{2u}{r\phi_r(r)F(r)} = \frac{u}{2mr}\ , \end{aligned} \tag{11}$$

where we have used the identity $\phi_r(r)F(r) = 4m$ in the last line. Similarly, we have that

$$\operatorname{div}_{\mathbb{S}^2(r)}(-\operatorname{grad} v) = -\frac{2\, r_u}{rF(r)} = -\frac{2v}{r\phi_r(r)F(r)} = -\frac{v}{2mr}. \tag{12}$$

The black hole region corresponds to $u < 0, v > 0$, in which case both divergences are *negative* and $\mathbb{S}^2(r)$ is what is called a *trapped surface; the remarkable thing here is that even the "outgoing" light rays are actually going inwards.* To paraphrase O'Neill: "These null geodesics have a better chance to escape than any particle starting inside $\mathbb{S}^2(r)$, hence their failure to do so looks like a good warning of the ensuing singularity" ([O'Neill, p. 434]).

3. "Coincidentally," *no* particle, whether material or light like, can escape from the black hole region of Kruskal spacetime (see [O'Neill, p. 392]). *Penrose essentially equated these two phenomena — trapped surfaces and "regions of no escape" — and this allowed him to characterize "regions of no escape" for arbitrary spacetimes (i.e., without reference to coordinates), because the condition of convergence/divergence is coordinate independent and makes sense for any spacelike submanifold.*

   In conclusion: Penrose replaced the statement "Let there be a 'region of no escape' in a spacetime" with the mathematically rigorous statement "Let a spacetime have a trapped surface."

4. *Trapped surface (Definition 1).* A trapped surface is a closed codimension 2 spacelike submanifold of a spacetime, both of whose future-pointing null normal vector fields are *converging.*

   This was the original definition of a trapped surface in [Penrose]. Note that we are taking our null normal vector fields to be *smooth* here; this is certainly always true locally, using adapted coordinate charts. We now recast this definition using the *mean curvature vector field* of a submanifold.

5. *Trapped surfaces via mean curvature.* The best way to characterize the condition of a trapped surface is via a submanifold's mean curvature vector field. (Recall that given a k-submanifold $P \subset (M, g)$ with second fundamental form tensor $I\!I$ and a slice coordinate chart $(\mathcal{U}, (x^i))$, the *mean curvature vector field* on $P \cap \mathcal{U}$ is given by

$$H = \frac{1}{k}\sum_{i=1}^{k} g^{ij} I\!I(\partial_i, \partial_j). \tag{13}$$

   To compute $I\!I(\partial_i, \partial_j)$, note that if the remaining vectors $\partial_{k+1}, \ldots, \partial_n$ are normal to P, then

$$I\!I(\partial_i, \partial_j) = \sum_{r=k+1}^{n} \widetilde{\Gamma}^r_{ij}\, \partial_r, \tag{14}$$

where $\widetilde{\Gamma}^r_{ij}$ are Christoffel symbols on the spacetime M (see [O'Neill, p. 101, 123]).) In terms of H, we have the following characterization of "trapped" submanifold. Namely, given a spacelike codimension ≥ 2 submanifold $P \subset M$ with mean curvature vector field H, the following are equivalent:

(a) $\langle H, v\rangle > 0$ for all future-pointing *null* vectors v normal to P,

(b) $\langle H, w\rangle > 0$ for all future-pointing *causal* vectors w normal to P,

(c) H is past-pointing timelike.

(Proof: (c) $\Rightarrow$ (b): Let $p \in P$ and $w_p \in (T_pP)^\perp$ a future-pointing causal vector. Now in general, two causal vectors H_p and w_p are in the same causal cone $\Leftrightarrow$ either $\langle H_p, w_p\rangle < 0$ or H_p and w_p are both null with $H_p = aw_p, a > 0$. Since H_p is past-pointing timelike, clearly we must have $\langle H_p, w_p\rangle \geq 0$, but $\langle H_p, w_p\rangle = 0$ cannot happen because only null causal vectors can be orthogonal (see [O'Neill, p. 155, Exercises 2(b),3(a)]). So it must be the case that $\langle H_p, w_p\rangle > 0$; (b) $\Rightarrow$ (a): Obvious; (a) $\Rightarrow$ (c): H_p certainly can't be null or else (a) would be contradicted. Can H_p be spacelike? Suppose it were; then our Lorentz vector space decomposes into $H_p \oplus H_p^\perp$, where $H_p^\perp \subset T_pM$ is an $(n-1)$-dimensional *Lorentz* vector space. But this means that any future-pointing null vector $v_p \in H_p^\perp$ satisfies $\langle H_p, v_p\rangle = 0$, a contradiction. Hence H_p must be timelike (and obviously past-pointing).) With this result, we make the following definition:

6. *Trapped surface (Definition 2).* A spacelike submanifold is a *trapped submanifold* if its mean curvature vector field is past-pointing timelike. If this submanifold has dimension two, then we say it is a *"trapped surface."* (This mathematical definition is crucial: we can now look at any spacelike submanifold, even a hypersurface, and ask whether it is trapped.)

   One question remains: for codimension 2 submanifolds like $\mathbb{S}^2(r)$ in Kruskal spacetime, are Definitions 1 and 2 equivalent? *Yes:* for codimension 2 the the two linearly independent null vectors at each point constitute a basis for the normal space, hence by eqn. (14) H can be expressed solely with respect to them. From there it is not difficult to show that H will be past-pointing timelike $\Leftrightarrow$ both future-pointing null normal vector fields are converging (Proof: **(FINISH!)**). *We will use Definition 2 to describe trapped surfaces.* Here's how it plays out for spheres $\mathbb{S}^2(r)$ in Kruskal spacetime.

7. *Trapped surfaces in Kruskal spacetime, this time via mean curvature.* Let's use eqn. (13) and eqn. (14) to compute the mean curvature vector field of a t-constant sphere $\mathbb{S}^2(r)$:

$$\begin{aligned} H &= \frac{1}{2}\sum_{i=\vartheta,\varphi} g^{ij}\,II(\partial_i,\partial_j) \\ &= \frac{1}{2}\left[\frac{II(\partial_\vartheta,\partial_\vartheta)}{r^2} + \frac{II(\partial_\varphi,\partial_\varphi)}{r^2\sin^2\vartheta}\right] \\ &= \frac{1}{2}\left[\frac{\widetilde{\Gamma}^u_{\vartheta\vartheta}\,\partial_u + \widetilde{\Gamma}^v_{\vartheta\vartheta}\,\partial_v}{r^2} + \frac{\widetilde{\Gamma}^u_{\varphi\varphi}\,\partial_u + \widetilde{\Gamma}^v_{\varphi\varphi}\,\partial_v}{r^2\sin^2\vartheta}\right] \\ &= -\frac{2\,r_v}{rF(r)}\partial_u - \frac{2\,r_u}{rF(r)}\partial_v \\ &= -\frac{u}{2mr}\partial_u - \frac{v}{2mr}\partial_v, \end{aligned}$$

   where the necessary computations were already carried out in eqns. (11) and (12) above. (Note that $H = -\mathrm{div}_{\mathbb{S}^2(r)}(\mathrm{grad}\,u)\,\partial_u - \mathrm{div}_{\mathbb{S}^2(r)}(\mathrm{grad}\,v)\,\partial_v$.) From here it's easy to verify that H is past-pointing timelike in the $u < 0, v > 0$ region of Kruskal spacetime:

$$\begin{aligned} \langle H, H\rangle_{\tilde{g}} = \frac{uv}{mr}\,\tilde{g}_{uv} < 0 \qquad &\Rightarrow \qquad H \text{ timelike,} \\ \langle H, \partial_v - \partial_u\rangle_{\tilde{g}} = -\frac{u}{2mr}\,\tilde{g}_{uv} + \frac{v}{2mr}\,\tilde{g}_{uv} > 0 \qquad &\Rightarrow \qquad H \text{ past-pointing.} \end{aligned}$$

This means that for any future-pointing causal vector v normal to $\mathbb{S}^2(r)$, we must have $\langle H, v\rangle_{\tilde{g}} > 0$. To see the significance of this, let $\alpha(s) = (r(s), t(s), \vartheta(s), \varphi(s))$ be a future-pointing causal curve in the $r < 2m$ region (we're now using interior Schwarzschild coordinates). Since $\operatorname{grad} r = \frac{u}{4m}\,\partial_u + \frac{v}{4m}\,\partial_v$ (via eqns. (11) and (12), noting that $\operatorname{grad} r = \tilde{g}^{uv}(r_u + r_v)$), we have that $H = -(1/r)\operatorname{grad} r$, and so

$$\langle H, \alpha'\rangle_{\tilde{g}} = -\frac{1}{r}\langle \operatorname{grad} r, \alpha'\rangle_{\tilde{g}} = -\frac{1}{r}\frac{dr}{ds} > 0 \quad \Leftrightarrow \quad \frac{dr}{ds} < 0. \tag{15}$$

Hence our "trapped surface" condition (the surface here being $\mathbb{S}^2(r)$) is *equivalent* (in Kruskal spacetime) to the "region of no return" condition, the latter expressed in the fact that the r coordinate of α is *decreasing;* in other words, that α has no choice but to move toward the singularity at $r = 0$ — and as we know, it will necessarily get there at a *finite* value of its parameter (see [O'Neill, p. 392]), so in particular *if it's a geodesic then it can't be complete.* This is the motivation behind Penrose's notion of a "singular spacetime." *Penrose's idea: let's define a "singular spacetime" as one that contains an incomplete timelike or null geodesic.*

8. *Trapped subsets.* Given any subset $P \subset M$, consider $E^+(P) := J^+(P) - I^+(P)$. Here are some properties of $E^+(P)$:

   - $E^+(P)$ *is achronal:* $E^+(P) \subset J^+(P) \Rightarrow I^+(E^+(P)) \subset I^+(J^+(P)) = I^+(P) \Rightarrow E^+(P) \cap I^+(E^+(P)) = \varnothing \Rightarrow E^+(P)$ achronal. (We'll get achronality again below.)
   - $P \subset E^+(P) \Leftrightarrow P$ *is achronal:* ($\Rightarrow$) Obvious, as subsets of achronal sets are achronal; ($\Leftarrow$) P achronal $\Rightarrow P \cap I^+(P) = \varnothing$, so $P \subset J^+(P) - I^+(P) = E^+(P)$ (recall the definition of causal future: $J^+(P) = \{q \ : \ p < q \text{ or } p = q \text{ for some } p \in P\}$).
   - *If P is compact, then $E^+(P)$ is a closed achronal topological hypersurface:* P compact $\Rightarrow J^+(P)$ closed (see [O'Neill, p. 438, Exercise 4]) which, together with the fact that $\operatorname{int} J^+(P) = I^+(P) \Rightarrow E^+(P) = \operatorname{bd} J^+(P)$ is the boundary of future set (see [O'Neill, p. 415]).
   - *When P is a spacelike submanifold, $E^+(P)$ is generated by conjugate-free P-normal null geodesics:* given $q \in E^+(P)$, let α be a causal curve from P to q; a variational result (see [O'Neill, p. 298]) tells us that there is a timelike curve from P to q arbitrarily near α *unless* α is a P-normal null geodesic without conjugate points before q. But if $q \in J^+(P) - I^+(P)$ then there can be *no* timelike curve from p to q, else $q \in I^+(P)$, a contradiction.

9. *Trapped surface $P \Rightarrow$ trapped (i.e., compact)* $\operatorname{bd} J^+(P)$. *(This is the key step in Penrose's proof.)*

   - We begin with our first condition:

     Let P be a compact, achronal trapped surface. (**C1**)

     Since P is a spacelike codimension 2 submanifold, each (two dimensional) normal space $(T_pP)^\perp$ is timelike, hence contains two linearly independent null vectors $N_\pm(p)$. Consider the subspace $\widetilde{P}$ of the normal bundle $NP = \coprod_{p\in P}(T_pP)^\perp \subset TM$ consisting of these,

     $$\widetilde{P} := \coprod_{p\in P}\{N_\pm(p)\} \subset NP.$$

   - $\widetilde{P}$ *is compact.* Start by considering the even smaller subspace $\widetilde{P}_+ := \coprod_{p\in P}\{N_+(p)\}$. Restricting the smooth submersion $\pi\colon NP \longrightarrow P$ to $\widetilde{P}_+$,

     $$\pi|_{\widetilde{P}}\colon \widetilde{P}_+ \longrightarrow \pi(\widetilde{P}_+) = P \qquad , \qquad N_+(p) \ \mapsto \ p,$$

     it follows that P compact $\Rightarrow \widetilde{P}_+$ compact (Proof: if $\{\widetilde{\mathcal{U}}_\alpha\}$ is an open cover of $\widetilde{P}_+$, then by the subspace topology each $\widetilde{\mathcal{U}}_\alpha = \widetilde{P}_+ \cap \mathcal{U}_\alpha$, some open set $\mathcal{U}_\alpha \subset NP$ (note that each $\mathcal{U}_\alpha$ must contain

the $N_+(p)$ of each normal space $(T_pP)^\perp$ that it meets). But any smooth submersion is an open map (because smooth submersions always admit local sections; see [LeeISM, p. 88]), so the open cover $\{\pi(\mathcal{U}_\alpha)\}$ of P has a finite subcover, say $\{\pi(\mathcal{U}_{\alpha_1}), \ldots, \pi(\mathcal{U}_{\alpha_r})\}$. Thus $\{\mathcal{U}_{\alpha_1}, \ldots, \mathcal{U}_{\alpha_r}\}$ must meet all the $(T_pP)^\perp$'s, hence contain all the $N_+(p)$'s. It follows that $\{\widetilde{\mathcal{U}}_{\alpha_1}, \ldots, \widetilde{\mathcal{U}}_{\alpha_r}\}$ is a finite subcover of $\widetilde{P}_+$). Applying the same argument to $\widetilde{P}_- := \coprod_{p\in P}\{N_-(p)\}$, we conlcude that $\widetilde{P}_-$ is also compact. Hence their union $\widetilde{P}_- \cup \widetilde{P}_+ = \widetilde{P}$ is compact as well.

- *All the null geodesics coming out of P have conjugate points.* For any codimension 2 submanifold P with mean curvature vector field H, if α is a null geodesic such that

$$\alpha'(0) \perp T_{\alpha(0)}P \qquad \text{and} \qquad \langle \alpha'(0), H_{\alpha(0)}\rangle_{\tilde{g}} := k_{\alpha(0)} > 0,$$

then provided that α is defined on $[0, 1/k_\alpha]$, there is necessarily a conjugate point of P along $\alpha|_{[0,1/k_\alpha]}$ (see [O'Neill, p. 292]). In order for this to happen, however, we need the following:

$$\mathrm{Ric}(v, v) \geq 0 \text{ for all null tangent vectors } v \text{ to } M. \qquad (\mathbf{C2})$$

So henceforth we impose this condition on our spacetime. Now, back to our problem:

(a) Our P is trapped $\Rightarrow$ H is past-pointing timelike $\Rightarrow \langle \alpha'(0), H_{\alpha(0)}\rangle_{\tilde{g}} > 0$ *is always satisfied for any future-pointing null geodesic.* So far so good.

(b) To ensure that each future-pointing null geodesic α is actually *defined* on its corresponding interval $[0, 1/k_\alpha]$, it is sufficient to assume that

$$M \text{ is future null geodesically complete.} \qquad (\mathbf{C3})$$

- *And we can find a "common interval" for all of these conjugate points.* The problem here is that we have infinitely many null normals: $N_\pm(p)$ at each $p \in P$. This means that the corresponding null geodesics $\gamma_{N_\pm}$ each have their own "$[0, 1/k_{\gamma_\pm}]$" parameter interval in which their conjugate points are contained; can we find one *finite* interval $[0, b]$ that contains all of these intervals, or are the $[0, 1/k_{\gamma_\pm}]$'s unbounded? *We can, and in fact that's where the compactness of $\widetilde{P}$ comes into play.* Define

$$k|_{\widetilde{P}} \colon \widetilde{P} \subset NP \longrightarrow \mathbb{R} \qquad , \qquad N_\pm(p) \;\mapsto\; \langle N_\pm(p), H_p\rangle_{\tilde{g}} := k_{p_\pm}.$$

This is a continuous map (because it's the restriction of the smooth map $k \colon NP \longrightarrow \mathbb{R}$, as can be seen by using an adapted coordinate chart for P). *Because $\widetilde{P}$ is compact,* the image $k|_{\widetilde{P}}(\widetilde{P}) \subset \mathbb{R}$ is bounded and realizes its extrema. Thus there is a *smallest* value for $k|_{\widetilde{P}}$, which we denote $1/b$, and hence a *largest finite* interval $[0, b]$. *So at the end of the day, we see that because P is compact, all future-pointing null geodesics $\gamma_{N_\pm}$ with normal tangents $N_\pm(p)$ have a conjugate point somewhere in the interval $[0, b]$.*

- *And this, finally, forces $E^+(P)$ to be compact.* We proceed in steps:

(a) First, consider any null geodesic β normal to P, so that $\beta'(0) \in \mathrm{span}(N_+(\beta(0)), N_-(\beta(0)))$. Since *no* linear combination $aN_+(\beta(0))+bN_-(\beta(0))$ will result in a null vector unless precisely one of the coefficients is zero, *all null geodesics normal to P must be (reparametrizations of) $\gamma_{N_\pm}$'s.* This means that the set $E^+(P)$, being generated by conjugate-free P-normal null geodesics, is comprised precisely *of all points along γ_{N_+} or γ_{N_-} before any conjugate points are reached.* (Below, we'll just write $\gamma_{N_\pm}$ to refer to any such geodesic.)

(b) Thus $q \in E^+(P) \Rightarrow q = \gamma_{N_\pm}(s)$ *and* $s \in [0, b]$ (because $\gamma_{N_\pm}$ necessarily has a conjugate point in $[0, b]$, but the portion *up to* q is conjugate-free, which means that $s \in [0, b]$).

(c) Define $K = \{sN_\pm \;:\; N_\pm \in \widetilde{P} \text{ and } 0 \leq s \leq b\}$. This is a subset of $\mathcal{E} \subset TM$, the domain of the exponential map. Indeed,

$$q \in E^+(P) \Rightarrow q = \gamma_{N_\pm}(s) = \gamma_{sN_\pm}(1) \Rightarrow q \in \exp(K) \Rightarrow E^+(P) \subset \exp(K).$$

Moreover, $\widetilde{P}$ compact $\Rightarrow$ K compact $\Rightarrow$ $\exp(K)$ compact. Does this force the subset $E^+(P)$ to be compact as well? *Yes:* given a sequence $\{\tilde{q}_n\} \subset E^+(P)$, we have

$$\begin{aligned} \{\tilde{q}_n\} \subset E^+(P) \subset \underbrace{\exp(K)}_{\text{compact}} \quad &\Rightarrow \quad \text{conv. subseq. } \{\tilde{q}_{n_j}\} \to \tilde{q} \in \exp(K) \\ &\Rightarrow \quad \tilde{q} = \gamma_{sN_\pm}(1) = \gamma_{N_\pm}(s), \text{ some } sN_- \text{ or } sN_+ \in K \\ &\Rightarrow \quad \tilde{q} \in J^+(P). \end{aligned}$$

Now, if $\tilde{q} \in I^+(P)$ then because $I^+(P)$ is an open set we must have some $\tilde{q}_{n_j} \in I^+(P)$, which of course can't happen because $\{\tilde{q}_n\} \subset E^+(P)$. Hence $\tilde{q} \in J^+(P) - I^+(P)$ and we're done.

In words: *the light rays emanating from a compact achronal trapped surface cannot go very far!* $\square$

10. **Theorem [Penrose]:** A spacetime (M, g) cannot simultaneously satisfy the following four properties:

(C0) (M, g) contains a *noncompact* Cauchy hypersurface S.

(C1) (M, g) contains a compact, achronal *trapped surface* P.

(C2) $\mathrm{Ric}(v, v) \geq 0$ for all null tangent vectors v to (M, g).

(C3) (M, g) is future null geodesically complete.

Proof: the goal is show that if **(C1), (C2),** and **(C3)** hold, then $E^+(P)$ and S must be homeomorphic.

- M has a Cauchy hypersurface $\Rightarrow$ M is globally hyperbolic $\Rightarrow$ all $J^\pm(p)$'s closed $\Rightarrow$ $J^+(K)$ closed whenever K is compact (exercise!) $\Rightarrow$ $J^+(P)$ is closed $\Rightarrow$ $E^+(P) = \mathrm{bd}\, J^+(P)$ is (among other things) a topological 3-manifold, and *compact* by **(C1), (C2),** and **(C3)**.
- The achronality of $E^+(P)$ is important, because when we restrict the retraction given by eqn. (9) to $r|_{E^+(P)} \colon E^+(P) \longrightarrow S$, it will be *injective:* if two points $p, q \in E^+(P)$ "flow" to the same point in S, then by the uniqueness of integral curves p and q must lie on *one* integral curve, which is impossible because then one timelike curve would intersect $E^+(P)$ twice, thus violating its achronality (by the way, P is achronal $\Leftrightarrow$ $P \subset E^+(P)$ (exercise!)).
- Because $E^+(P)$ and S are topological manifolds of the same dimension (and neither has a boundary), the invariance of domain then says that $r|_{E^+(P)}$ is an open map. But $E^+(P)$ compact $\Rightarrow$ $r(E^+(P))$ compact $\Rightarrow$ $r(E^+(P))$ closed. Hence $r(E^+(P))$ is both open and closed in a connected set S, so we must have $r(E^+(P)) = S$. Thus $E^+(P)$ is homeomorphic to the Cauchy hypersurface S. But the former is compact and the latter is not. $\square$

In particular, if **(C0), (C1),** *and* **(C2)** *hold on* (M, g)*, then it is future null geodesically incomplete!* Notice that *no assumption of spherically symmetry has been made here,* and this is one of the great virtues of Penrose's result: only a trapped surface is required to exist in order for a singularity to be inevitable. Notice also that the requirement of a noncompact Cauchy hypersurface is not too stringent, as one perhaps expects this to be the case for asymptotically flat spacetimes. Finally, notice that no mention of the inextendibility of the spacetime is made: this is something that must be verified independently.

Lecture IV: Gannon's singularity theorem

1. Let (M, g) be a spacetime with a *nonsimply connected* Cauchy hypersurface S.

   - Before we begin, let us collect several important properties regarding covering maps:
     (a) *Definition:* a *smooth covering map* $\pi\colon \widetilde{M} \longrightarrow M$ between two smooth manifolds $M, \widetilde{M}$ is a smooth surjective map with the following property: for each $p \in M$ there is a neighborhood $\mathcal{U}$ of p such that each component of $\pi^{-1}(\mathcal{U})$ is mapped diffeomorphically to $\mathcal{U}$ by π (for a connected topological manifold, path components = connected components). If $\widetilde{M}$ is simply connected, then $\pi\colon \widetilde{M} \longrightarrow M$ is called the *universal covering of* M. The components of $\pi^{-1}(\mathcal{U})$ are often called the *slices of* $\mathcal{U}$.
     (b) *"Topological" vs. "smooth" covering maps:* the former is a *continuous* surjective map whose disjoint slices upstairs are *homeomorphic* to their corresponding evenly covered neighborhoods downstairs; the latter is a *smooth* surjective map whose disjoint slices upstairs are *diffeomorphic* to their corresponding evenly covered neighborhoods downstairs. In particular, a smooth covering map is more than just a topological covering map that happens to be smooth (see [LeeISM, p. 91] for more on this distinction). Given that Cauchy hypersurfaces are in general *not* smooth, this is a distinction we should be mindful of.
     (c) *All topological (smooth) manifolds X have topological (smooth) universal covering maps:* this is a covering map $\pi\colon \widetilde{X} \longrightarrow X$ in which $\widetilde{X}$ is simply connected, that is, $\pi_1(\widetilde{X}) = 0$. If X is a connected smooth manifold, then its universal cover $\widetilde{X}$ will be a smooth manifold and π a smooth covering map (see [LeeISM, p. 94]). All universal covers are *normal,* meaning that $\pi_*(\pi_1(\widetilde{X})) = 0 \subset \pi_1(X)$ is a normal subgroup (trivially in this case); this is a necessary and sufficient condition to ensure that the deck transformation group $\mathrm{Aut}_\pi(\widetilde{X})$ acts transitively on the fibers of X (see [LeeISM, p. 163]).
     (d) *All covering maps have a lifting criterion:* given a topological covering map $\pi\colon C \longrightarrow X$ and a continuous map $h\colon Y \longrightarrow X$ with Y a connected topological manifold, there is a unique continuous lift $\tilde{h}\colon Y \longrightarrow C$

     $$\begin{array}{ccc} & & (C, c_0) \\ & \overset{\tilde{h}}{\nearrow} & \big\downarrow \pi \\ (Y, y_0) & \xrightarrow{\ h\ } & (X, x_0) \end{array}$$

     satisfying $h(y_0) = c_0$ and making the diagram commute $\Leftrightarrow h_*(\pi_1(Y, y_0)) \subset \pi_*(\pi_1(C, c_0)) \subset \pi_1(X, x_0)$ (see [LeeISM, p. 616]). If "topological" is replaced by "smooth" everywhere above and h is a curve (so $Y \subset \mathbb{R}$), then the lift $\tilde{h}$ will be a smooth curve (see [O'Neill, p. 444]).
     (e) Let (N, g) be a Lorentzian manifold. Given the smooth map $F\colon M \longrightarrow N$, F^*g is a Lorentzian metric on $F \Leftrightarrow F$ is a smooth immersion (see [LeeISM, p. 331]). Since any smooth covering map $C \longrightarrow X$ is a local diffeomorphism and hence an immersion, *C can always be made into a Lorentzian manifold provided that X is one.*
   - Now we begin. Recall that M retracts onto the topological 3-manifold S and is homeomorphic to $\mathbb{R} \times S$, so in particular S is connected and $\pi_1(S) \cong \pi_1(M)$. Let $r\colon M \longrightarrow S$ denote this (continuous) retraction. Also, let $\pi\colon \widetilde{M} \longrightarrow M$ denote the smooth universal cover of M. Note that the restriction $\pi|_S\colon \pi^{-1}(S) \longrightarrow S$ is a (topological) covering map (any restriction of a topological covering map is a topological covering map).
   - *$\widetilde{M}$ is a spacetime in its own right:* being a universal cover, $\widetilde{M}$ is connected. Moreover, g lifts to a Lorentzian metric $\tilde{g} = \pi^*g$ on $\widetilde{M}$, *hence $(\widetilde{M}, \tilde{g})$ and (M, g) are locally isometric.* As a consequence, suppose, for example, that $\tilde{\alpha}$ is a timelike curve in $\widetilde{M}$; then $\tilde{\alpha}$ projects to a timelike

curve $\pi \circ \tilde{\alpha} := \alpha$ in M:

$$\begin{aligned}\langle \alpha'(s), \alpha'(s) \rangle_g &= \langle d(\pi \circ \tilde{\alpha})|_{\alpha(s)}(d/ds), d(\pi \circ \tilde{\alpha})|_{\alpha(s)}(d/ds) \rangle_g \\ &= \pi^* \langle d\tilde{\alpha}|_{\tilde{\alpha}(s)}(d/ds), d\tilde{\alpha}|_{\tilde{\alpha}(s)}(d/ds) \rangle_g \\ &= \langle \tilde{\alpha}'(s), \tilde{\alpha}'(s) \rangle_{\tilde{g}} < 0.\end{aligned}$$

Hence the causal structure of a curve in $\widetilde{M}$ is preserved when it descends to M, as is its possible inextendibility, because π is a local isometry. Moreover, $(\widetilde{M}, \tilde{g})$ can be time-oriented via the time orientation $T \in \mathfrak{X}(M)$ (Proof: for any $p \in M$, let $\pi|_{\widetilde{\mathcal{U}}} \colon \widetilde{\mathcal{U}} \longrightarrow \mathcal{U}$ denote the diffeomorphism corresponding to a slice $\widetilde{\mathcal{U}}$ of the evenly covered neighborhood $\mathcal{U}$ of p. Then $X|_{\mathcal{U}}$ is π-related to the smooth timelike vector field $\widetilde{X}|_{\widetilde{\mathcal{U}}} \in \mathfrak{X}(\widetilde{\mathcal{U}})$ defined by $\widetilde{T}_{\tilde{p}} := (d\pi)^{-1}(T_{\pi(\tilde{p})}) \in T_{\tilde{p}}\widetilde{M}$. These local time orientations will then piece together into a global time orientation $\widetilde{T} \in \mathfrak{X}(\widetilde{M})$; see [LeeISM, p. 183]). It follows that future-pointing curves in $\widetilde{M}$ will descend to future-pointing ones in M. *Moreover, inextendible integral curves of $\widetilde{T}$ project down to inextendible integral curves of T.* Indeed, if $\tilde{\theta}^{(\tilde{p})} \colon \widetilde{\mathcal{D}}^{(\tilde{p})} \longrightarrow \widetilde{M}$ is the maximal integral curve of $\widetilde{T}$ starting at $\tilde{p} \in \widetilde{M}$, then $\pi \circ \tilde{\theta}^{(\tilde{p})} \colon \widetilde{\mathcal{D}}^{(\tilde{p})} \longrightarrow M$ is the maximal integral curve $\theta^{(p)} \colon \mathcal{D}^{(p)} \longrightarrow \widetilde{M}$ of T starting at $\pi(\tilde{p}) = p \in M$, since

$$d(\pi \circ \tilde{\theta}^{(\tilde{p})}) \left(\left. \frac{d}{dt} \right|_t \right) = d\pi|_{\tilde{\theta}^{(\tilde{p})}(t)}(\widetilde{T}_{\tilde{\theta}^{(\tilde{p})}(t)}) = T_{\pi \circ \tilde{\theta}^{(\tilde{p})}(t)} \Rightarrow \pi \circ \tilde{\theta}^{(\tilde{p})}(t) = \theta^{(p)}(t). \tag{16}$$

By uniqueness of solutions to ODEs with initial conditions, we must have $\pi \circ \tilde{\theta}^{(\tilde{p})} \equiv \theta^{(p)}$ wherever their domains coincide. And in fact we must have $\widetilde{\mathcal{D}}^{(\tilde{p})} = \mathcal{D}^{(p)}$ because π is a local diffeomorphism, so it preserves inextendibility.

- $\pi^{-1}(S)$ *is a Cauchy hypersurface in $\widetilde{M}$:* any inextendible timelike curve $\tilde{\alpha}$ in $\widetilde{M}$ projects to an inextendible timelike curve $\pi \circ \tilde{\alpha} := \alpha$ in M; since α must meet S exactly once, $\tilde{\alpha}$ meets $\pi^{-1}(S)$ exactly once (if $\tilde{\alpha}$ met $\pi^{-1}(S)$ more than once, then S would not be achronal, a contradiction). *In particular, $\widetilde{M}$ retracts onto $\pi^{-1}(S)$ and is homeomorphic to $\mathbb{R} \times \pi^{-1}(S)$, so the $\pi^{-1}(S)$ is simply connected and $\pi|_S$ is the universal cover of S.* Denoting this retraction by $\tilde{r}$, the discussion leading to eqn. (16) then implies that the following diagram commutes, where "u. c." denotes "universal cover":

$$\begin{array}{ccc} \overbrace{\pi^{-1}(S)}^{\text{u. c.}} & \xleftarrow{\tilde{r}} & \overbrace{\widetilde{M}}^{\text{u. c.}} \\ \Big\downarrow{\scriptstyle \pi|_S} & & \Big\downarrow{\scriptstyle \pi} \\ S & \xleftarrow{r} & M. \end{array}$$

Here's why: given any $\tilde{p} \in \widetilde{M}$, $\tilde{r}(\tilde{p})$ is by definition the unique point in $\pi^{-1}(S)$ hit by the integral curve $\tilde{\theta}^{(\tilde{p})}(t)$; let us denote this point by $\tilde{\theta}^{(\tilde{p})}(t_1)$. Then we have

$$\tilde{p} \;\mapsto\; \tilde{r}(\tilde{p}) = \tilde{\theta}^{(\tilde{p})}(t_1) \;\mapsto\; \underbrace{\pi|_S \circ \tilde{\theta}^{(\tilde{p})}(t_1) = \theta^{(\pi(\tilde{p}))}(t_1)}_{\text{by eqn. (16)}} = r \circ \pi(\tilde{p}).$$

(By the way, because this diagram commutes, $\tilde{r}$ is in fact the unique lift "$\tilde{h}$" of the map $r \circ \pi$, provided we stipulate that $\tilde{h}(\tilde{p}) = \tilde{p}$ for *any* $\tilde{p} \in \pi^{-1}(S)$:

$$\begin{array}{ccccc} & & & & \pi^{-1}(S) \\ & & \overset{\tilde{h}}{\nearrow} & & \Big\downarrow{\scriptstyle \pi|_S} \\ \widetilde{M} & \xrightarrow{\pi} & M & \xrightarrow{r} & S. \end{array}$$

In other words, since $\tilde{h}(\tilde{p}) = \tilde{p} = \tilde{r}(\tilde{p})$ and $\pi|_S \circ \tilde{h} = r \circ \pi = \pi|_S \circ \tilde{r}$, by uniqueness of lifts we *must* have $\tilde{h} \equiv \tilde{r}$. Now suppose we didn't know about $\tilde{r}$, and we only had $\tilde{h}$ to work with. Does it follow that $\tilde{h}$ is indeed a retraction? *Yes:* pick any other point $\tilde{q} \in \pi^{-1}(S)$ and consider a path $\tilde{p} \xrightarrow{\tilde{\alpha}} \tilde{q}$ as well as $\tilde{p} \xrightarrow{\tilde{h}\circ\tilde{\alpha}} \tilde{h}(\tilde{q})$ (remember that $\pi^{-1}(S)$ is a connected topological manifold, hence both $\tilde{\alpha}$ and $\tilde{h} \circ \tilde{\alpha}$ are paths contained in $\pi^{-1}(S)$). Letting $\pi|_S \circ \tilde{\alpha} := \alpha$, we have

$$\pi|_S \circ (\tilde{h} \circ \tilde{\alpha}) = (r \circ \pi) \circ \tilde{\alpha} = \underbrace{r \circ \alpha = \alpha}_{r|_S \equiv \mathrm{id}} \quad \Rightarrow \quad \underbrace{\tilde{h} \circ \tilde{\alpha} \equiv \tilde{\alpha}}_{\text{uniqueness of lifts}} \quad \Rightarrow \tilde{h}(\tilde{q}) = \tilde{q}. \tag{17}$$

Hence $\tilde{h}$ is indeed a retraction, because r was. Moreover, the homomorphism $\tilde{\iota}_*$ induced by inclusion $\tilde{\iota}\colon \pi^{-1}(S) \hookrightarrow \widetilde{M}$ is injective, which implies that $\pi^{-1}(S)$ is simply connected and $\pi|_S\colon \pi^{-1}(S) \longrightarrow S$ is the universal cover of S).

[Gannon] proves that, subject to some conditions on S, nontrivial topology $\pi_1(S) \neq 0$ guarantees the existence of an incomplete null geodesic in the universal cover $\widetilde{M}$, which will then project down to an incomplete null geodesic in M. In particular, there is no assumption that M should contain a trapped surface, although the null energy condition will be assumed.

2. *Gannon's regularity conditions:*

   (a) $S = \bigcup_{i=1}^{\infty} W_i$ is a spacelike Cauchy hypersurface, where the W_i's are a nested sequence

   $$W_1 \subset W_2 \subset W_3 \subset \cdots$$

   and each W_i is a compact smooth 3-manifold whose boundary ∂W_i is homeomorphic to $\mathbb{S}^2$.

   (b) $S - \operatorname{int} W_i$ is homemorphic to $\partial W_i \times [0, \infty)$. In particular, S retracts onto any W_i.

   (c) The "inward" directed null geodesics normal to ∂W_i are converging. (This is what we expect for spheres "outsde" the black hole; see, for example, eqns. (11) and (12) when $u > 0, v > 0$.)

   Except for the compactness condition, these conditions are satisfied by all asymptotically flat hypersurfaces in asymptotically flat spacetimes (see [Wald, p. 241]).

3. Since S is nonsimply connected and retracts onto each W_i by condition (b), some $W_i := W$ must be nonsimply connected. Then the restriction $\pi|_{\partial W}\colon \pi^{-1}(\partial W) \longrightarrow \partial W$ is a topological covering map, and as we now show, *each path component $\widetilde{P} \subset \pi^{-1}(\partial W)$ is diffeomorphic to ∂W.* Indeed, let $\pi_{\widetilde{P}}\colon \widetilde{P} \longrightarrow \partial W$ denote the restriction of the domain of $\pi|_{\partial W}$ to a path component $\widetilde{P}$. We start by showing that $\pi_{\widetilde{P}}$ is a homeomorphism:

   - $\pi_{\widetilde{P}}\colon \widetilde{P} \longrightarrow \partial W$ *is surjective:* pick any "basepoint" $\tilde{x} \in \widetilde{P}$ and set $\pi(\tilde{x}) = x \in \partial W$. Then for any $y \in \partial W$, there is a path $x \xrightarrow{\alpha} y$ (∂W is homeomorphic to $\mathbb{S}^2$, hence is path connected) which lifts via the covering $\pi|_{\partial W}$ to a path $\tilde{x} \xrightarrow{\tilde{\alpha}} \tilde{y}$. We would like to show that $\tilde{y}$ has to be in $\widetilde{P}$, for then $\pi_{\widetilde{P}}(\tilde{y}) = y$ and $\pi_{\widetilde{P}}$ would be surjective. But this is easy: since $\tilde{x} \in \widetilde{P}$, the path $\tilde{\alpha}$ must *stay* in the path component $\widetilde{P}$, so we're done:

   $$\begin{array}{ccc} \tilde{x} \xrightarrow{\tilde{\alpha}} \tilde{y} \subset \widetilde{P} \subset \pi^{-1}(\partial W) \\ \Big\downarrow{\scriptstyle \pi|_{\partial W}} \\ \pi(\tilde{x}) = x \xrightarrow{\alpha} y \subset \partial W. \end{array}$$

   - $\pi_{\widetilde{P}}\colon \widetilde{P} \longrightarrow \partial W$ *is a topological covering map, hence a local homeomorphism:* just make sure those neighborhoods $\mathcal{U}$ evenly covered by $\pi|_{\partial W}$ are path-connected (by possibly shrinking $\mathcal{U}$; recall that

all manifolds are locally path-connected, meaning they have a basis of path-connected open sets). Then any slice $\widetilde{\mathcal{U}}$ upstairs is also path-connected and must therefore be entirely contained in a path component of $\pi^{-1}(\partial W)$. The one contained in $\widetilde{P}$ will then serve as the slice for $\pi_{\widetilde{P}}\colon \widetilde{P} \longrightarrow \partial W$.

- $\pi_{\widetilde{P}}\colon \widetilde{P} \longrightarrow \partial W$ *is injective:* being homeomorphic to $\mathbb{S}^2$, ∂W is simply connected and so serves as its own universal cover $\mathrm{id}\colon \partial W \longrightarrow \partial W$. This map then lifts to a unique map $\tilde{h}\colon \partial W \longrightarrow \widetilde{P}$ making the following diagram commute:

$$\begin{array}{ccc} & & \widetilde{P} \\ & \overset{\tilde{h}}{\nearrow} & \big\downarrow{\scriptstyle \pi_{\widetilde{P}}} \\ \partial W & \xrightarrow{\mathrm{id}} & \partial W. \end{array}$$

  Now suppose there exist $\tilde{x}_1, \tilde{x}_2 \in \widetilde{P}$ such that $\pi_{\widetilde{P}}(\tilde{x}_1) = \pi_{\widetilde{P}}(\tilde{x}_2) = x$, and let us suppose that $\tilde{h}(x) = \tilde{x}_1$ (there's a unique lift sending x to each point in the fiber $\pi_{\widetilde{P}}^{-1}(x)$). Consider a path $\tilde{x}_1 \xrightarrow{\tilde{\alpha}} \tilde{x}_2$; its image $\pi_{\widetilde{P}} \circ \tilde{\alpha} := \alpha$ is a *loop* at $x \in \partial W$, and $\tilde{\alpha}$ is of course its (unique) lift starting at $\tilde{x}_1$. We now show that $\tilde{\alpha}$ also has to be a loop, so that $\tilde{x}_1 = \tilde{x}_2$. Indeed,

$$\alpha = \underbrace{(\pi_{\widetilde{P}} \circ \tilde{h})}_{\mathrm{id}} \circ \alpha = \pi_{\widetilde{P}} \circ \underbrace{(\tilde{h} \circ \alpha)}_{\text{loop at } \tilde{x}_1} \;\Rightarrow\; \underbrace{\tilde{\alpha} = \tilde{h} \circ \alpha}_{\text{uniqueness of lifts}} \Rightarrow \tilde{x}_1 = \tilde{x}_2.$$

  Of course, a bijective local homeomorphism is a homeomorphism, so we're done: each path component $\widetilde{P} \subset \pi^{-1}(\partial W)$ is homeomorphic to ∂W (which is achronal). *But we can say more: since ∂W is a connected smooth manifold and $\pi_{\widetilde{P}}\colon \widetilde{P} \longrightarrow \partial W$ is a topological covering map, one can endow $\widetilde{P}$ with a smooth structure such that $\pi_{\widetilde{P}}\colon \widetilde{P} \longrightarrow \partial W$ is a smooth covering map, hence a diffeomorphism* (see [LeeISM, p. 92]). Moreover, since ∂W is homeomorphic to $\mathbb{S}^2$, it follows that *each component $\widetilde{P}$ is compact and simply connected, and by (c) of Gannon's regularity conditions, the "inward" directed null geodesics normal to $\widetilde{P}$ are converging* (because they are inherited from those normal to ∂W via the diffeomorphism $\pi_{\widetilde{P}}$). So we now have the following inclusions:

$$\begin{array}{ccccccc} \widetilde{P} & \hookrightarrow & \overbrace{\pi^{-1}(\partial W)}^{\partial\pi^{-1}(W)} & \hookrightarrow & \overbrace{\pi^{-1}(S)}^{\text{u. c.}} & \hookrightarrow & \overbrace{\widetilde{M}}^{\text{u. c.}} \\ & \searrow^{\approx} & \big\downarrow{\scriptstyle \pi|_{\partial W}} & & \big\downarrow{\scriptstyle \pi|_S} & & \big\downarrow{\scriptstyle \pi} \\ & & \partial W & \hookrightarrow & S & \hookrightarrow & M. \end{array}$$

  Note that we have inserted the fact that $\pi^{-1}(\partial W) = \partial\pi^{-1}(W)$. Now fix a path component $\widetilde{P} \subset \partial(\pi^{-1}(W))$; *since we are not assuming that the "outward" directed null geodesics normal to P are converging, it is not assumed that P is a trapped surface; hence the same is true for $\widetilde{P}$ upstairs.*

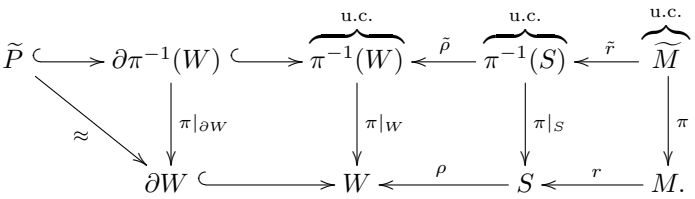
4. $\pi^{-1}(S)$ *(and hence $\widetilde{M}$) retracts onto $\pi^{-1}(W)$:* start by considering the topological covering map $\pi|_W\colon \pi^{-1}(W) \longrightarrow W$. By condition (b), S retracts onto W. Denoting this retraction by ρ, this implies that $\pi^{-1}(S)$ retracts onto $\pi^{-1}(W)$ (Proof: since $\pi^{-1}(S)$ is simply connected, the composite map $\rho \circ \pi|_S\colon \pi^{-1}(S) \longrightarrow W$ lifts to a unique map $\tilde{\rho}\colon \pi^{-1}(S) \longrightarrow \pi^{-1}(W)$ satisfying $\tilde{\rho}(\tilde{p}) = \tilde{p}$ for some $\tilde{p} \in \pi^{-1}(W)$, and making the following diagram commute:

$$\begin{array}{ccccc} & & & & \pi^{-1}(W) \\ & & \overset{\tilde{\rho}}{\nearrow} & & \big\downarrow{\scriptstyle \pi|_W} \\ \pi^{-1}(S) & \xrightarrow{\pi|_S} & S & \xrightarrow{\rho} & W. \end{array}$$

We need to check that $\tilde{\rho}$ is actually a retraction; in other words, that $\tilde{\rho}|_{\pi^{-1}(W)} \equiv \mathrm{id}$. To show this, we'll argue exactly as we did in eqn. (17) above. First, let's assume for the moment that $\pi^{-1}(W)$ is path connected. Under this assumption, pick any other point $\tilde{q} \in \pi^{-1}(W)$ and consider a path $\tilde{p} \xrightarrow{\tilde{\alpha}} \tilde{q}$ in $\pi^{-1}(W)$ as well as $\tilde{p} \xrightarrow{\tilde{\rho}\circ\tilde{\alpha}} \tilde{\rho}(\tilde{q})$, the latter automatically being in $\pi^{-1}(W)$. Then letting $\pi \circ \tilde{\alpha} := \alpha$, we have

$$\pi|_W \circ (\tilde{\rho} \circ \tilde{\alpha}) = (\rho \circ \pi|_S) \circ \tilde{\alpha} = \underbrace{\rho \circ \alpha = \alpha}_{\rho|_W \equiv \mathrm{id}} \quad \Rightarrow \quad \underbrace{\tilde{\rho} \circ \tilde{\alpha} \equiv \tilde{\alpha}}_{\text{uniqueness of lifts}} \quad \Rightarrow \tilde{\rho}(\tilde{q}) = \tilde{q}.$$

Hence $\tilde{\rho}$ is indeed a retraction, because ρ was. Now, if $\pi^{-1}(W)$ is *not* path connected, then our argument works for any one of its path components $\widetilde{L} \subset \pi^{-1}(W)$; in other words, we would restrict to the covering map $\pi|_{\widetilde{L}} \colon \widetilde{L} \longrightarrow W$ (this would be a covering map for exactly the same reason as $\pi_{\widetilde{P}}$ was above), in which case the lift $\tilde{\rho} \colon \pi^{-1}(S) \longrightarrow \widetilde{L}$ would be a retraction. Then for any $\tilde{p} \in \widetilde{L}$, we would of course have $\pi_1(\widetilde{L}, \tilde{p}) = \pi_1(\pi^{-1}(W), \tilde{p})$. For this reason, we'll assume henceforth that $\pi^{-1}(W)$ is path connected). Hence the homomorphism $\tilde{\iota}_*$ induced by inclusion $\tilde{\iota} \colon \pi^{-1}(W) \hookrightarrow \pi^{-1}(S)$ is injective, which implies that $\pi^{-1}(W)$ *is simply connected and* $\pi|_W \colon \pi^{-1}(W) \longrightarrow W$ *is the universal cover of* W. So if we put everything that we've gathered up to now together, the picture is the following commutative diagram:



5. *There is more than one path component* $\widetilde{P} \subset \partial\pi^{-1}(W)$*: this is where we use the fact that* S *is not simply connected and that* $\widetilde{M}$ *is the universal cover.* In particular, for any pair $\pi(\tilde{x}) = x \in \partial W$,

$$\begin{aligned} \#\text{ of path components of } \pi^{-1}(\partial W) &= |(\pi|_{\partial W})^{-1}(x)| \\ &= |\pi^{-1}(x)| \\ &= \text{index of } \underbrace{\pi_*(\pi_1(\widetilde{M}, \tilde{x}))}_{0} \subset \pi_1(M, x), \end{aligned}$$

which is therefore *greater than one* since $\pi_1(M, x) = \pi_1(S, x) \neq 0$ (since M is connected, the fiber $\pi^{-1}(x)$ is not just locally constant, but actually constant over all of M).

6. $\widetilde{P}$ *cannot bound any compact 3-manifold: this is where we use the fact that* $\widetilde{M}$ *retracts onto* $\pi^{-1}(W)$. In particular, no path component $\widetilde{P}$ is the (full) boundary $\partial\pi^{-1}(W)$ of $\pi^{-1}(W)$. But we can say more: $\widetilde{P}$ cannot be the boundary of *any compact* 3-manifold in $\pi^{-1}(S)$, and hence also in $\widetilde{M}$ since the latter retracts onto $\pi^{-1}(S)$ (Proof: suppose that $\widetilde{P} = \partial\widetilde{C}$ for some compact 3-manifold $\widetilde{C} \subset \widetilde{M}$, so in particular $\widetilde{P}$ is a compact 2-manifold without boundary, compact because any open cover of $\partial\widetilde{C}$ extends via $\operatorname{int}\widetilde{C}$ to an open cover of $\widetilde{C}$ (or even more trivially in our case, because we established that $\widetilde{P}$ is homeomorphic to the compact ∂W). Then $\widetilde{P}$ is in the zero homology class in $H_2(\widetilde{M})$. *Since* $\widetilde{M}$ *retracts onto* $\pi^{-1}(W)$ *via* $\tilde{\rho} \circ \tilde{r}$*,* the inclusion $\iota \colon \pi^{-1}(W) \hookrightarrow \widetilde{M}$ induces an injective map $\iota_* \colon H_2(\pi^{-1}(W)) \longrightarrow H_2(\widetilde{M})$, hence $\widetilde{P}$ must also be in the zero homology class in $H_2(\pi^{-1}(W))$. But this cannot happen if there is more than one boundary component in $\pi^{-1}(W)$).

7. Now fix a path component $\widetilde{P} \subset \partial\pi^{-1}(W)$ and focus attention on $\partial J^+(\widetilde{P}) \subset \widetilde{M}$. Being the boundary of a future set in the spacetime $\widetilde{M}$, $\partial J^+(\widetilde{P})$ is a closed achronal topological 3-manifold (see [O'Neill, p. 415]).

Moreover, $\mathrm{bd}\, J^+(\widetilde{P}) = J^+(\widetilde{P}) - I^+(\widetilde{P})$ is generated by $\widetilde{P}$-normal null geodesics (recall that $\widetilde{P}$ compact $\Rightarrow J^+(\widetilde{P})$ closed). By condition (c), the "inward" family of null geodesics is converging. As before, we're assuming the null energy condition on M, so $\mathrm{Ric}(v,v) \geq 0$ for all null vectors $v \in M \Rightarrow \widetilde{\mathrm{Ric}}(\tilde{v},\tilde{v}) \geq 0$ for all null vectors $\tilde{v} \in \widetilde{M}$. It is tempting at this point to argue that the mean curvature vector field $\widetilde{H}$ of $\widetilde{P}$ must satisfy $\langle \tilde{\alpha}'(0), \widetilde{H}_{\tilde{\alpha}(0)} \rangle_{\tilde{g}} > 0$ for all $\tilde{\alpha}$ in this "inward" family, so that each such $\tilde{\alpha}$ has a conjugate point somewhere within $\tilde{\alpha}|_{[0,1/k_{\tilde{\alpha}}]}$. The problem with this line of reasoning is that if $\langle \tilde{\alpha}'(0), \widetilde{H}_{\tilde{\alpha}(0)} \rangle_{\tilde{g}} > 0$ at each $\tilde{\alpha}(0) \in \widetilde{P}$, then $\widetilde{H}$ is past-pointing timelike, which would imply that $\widetilde{P}$ is a *trapped surface.* But we are taking care *not* to assume this, so we will have to argue for the existence of conjugate points in a different way. In fact the existence of conjugate points follows easily using the convergence property of the "inward" null vector field (see [Geroch2, p. 283]). Then just as before, these geodesics are no longer in $\partial J^+(\widetilde{P})$ after their first conjugate points.

8. Next, note that the "inward" family of ∂W-normal null geodesics *can never meet* the "outward" family (see [Gannon]); hence the same is true with respect to the two null vector fields normal to $\widetilde{P}$ upstairs. Assume now that each of the "inward" null geodesics normal to $\widetilde{P}$ is future complete. *Hence the proper submanifold determined by only the "inward" $\widetilde{P}$-normal null geodesics is compact with boundary $\widetilde{P}$.* But this would make $\widetilde{P}$ the (full) boundary of a compact 3-manifold $\widetilde{I}$, which as we noted above, is impossible. Hence the "inward" family of null geodesics normal to $\widetilde{P}$ *cannot* be complete: there must be at least one null geodesic from this "inward" family that is not defined up to its first conjugate point, so that $\widetilde{I}$ in fact never becomes a compact 3-manifold. Any such incomplete geodesic then descends downstairs to an incomplete null geodesic in M.

References

[BEE] Beem, John K., Ehrich, Paul E., and Easley, Kevin L., *Global Lorentzian Geometry,* second edition (Marcel Dekker, New York, 1996).

[Bernal-Sanchéz03] Bernal, Antonio M., and Sánchez, Miguel, "On smooth Cauchy hypersurfaces and Geroch's splitting theorem," *Comm. Math. Phys.* **243** (2003), 461 - 470.

[Bernal-Sanchéz07] Bernal, Antonio M., and Sánchez, Miguel, "Globally hyperbolic spacetimes can be defined as 'causal' instead of 'strongly causal'," *Class. Quant. Grav.* **24** (2007), 745 - 750.

[Chernov-Nemirovski] Chernov, Vladimir, and Nemirovski, Stefan, "Cosmic censorship of smooth structures," arXiv:1201.6070.

[Dieckmann] Dieckmann, J., "Volume functions in General Relativity," *Gen. Rel. Grav.* **20** (1988), 859 - 867.

[FSW] Friedman, John L., Schleich, Kristin, and Witt, Donald M., "Topological censorship," *Phys. Rev. Lett.* **71** (1993), 1486 -1489.

[Galloway] Galloway, Gregory J., "On the topology of the domain of outer communication," *Class. Quant. Grav.* **12** (1995), L99 - L101.

[Gannon] Gannon, Dennis, "Singularities in nonsimply connected space-times," *J. Math. Phys.* **16** (1975), 2364 - 2367.

[Geroch] Geroch, Robert, "Domain of Dependence," *J. Math. Phys.* **11** (1970), 437 - 449.

[Geroch2] Geroch, Robert, "Singularities," in *Relativity,* ed. Carmelli, M., Fickler, S., and Witten, L. (Plenum, New York, 1970).

[Lee] Lee, C. W., "A restriction on the topology of Cauchy surfaces in general relativity," *Comm. Math. Phys.* **51** (1976), 157 - 162.

[LeeRM] Lee, John M., *Riemannian Manifolds. An Introduction to Curvature* (Springer, New York, 1997).

[LeeISM] Lee, John M., *Introduction to Smooth Manifolds,* second edition (Springer, New York, 2012).

[O'Neill] O'Neill, Barrett, *Semi-Riemannian Geometry with Applications to Relativity* (Academic Press, New York, 1983).

[Penrose] Penrose, Roger, "Gravitational collapse and space-time singularities," *Phys. Rev. Lett.* **14** (1965), 57 - 59.

[Wald] Wald, Robert M., *General Relativity* (University of Chicago Press, Chicago, 1984).

[Wald-Iyer] Wald, Robert M., and Iyer, Vivek, "Trapped surfaces in the Schwarzschild geometry and cosmic censorship," *Phys. Rev. D* **44** (1991), 3719 - 3722.

[Witt] Witt, Donald M., "Vacuum space-times that admit no maximal slice," *Phys. Rev. Lett.* **57** (1986), 1386 - 1388.